Exome and Genome Sequencing

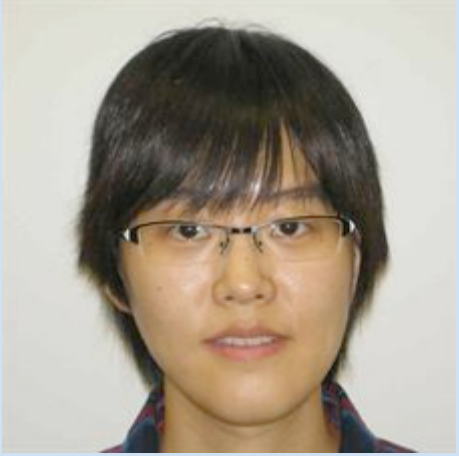
Gholson J. Lyon, M.D. Ph.D.

CSH Cold Spring Harbor Laboratory
Stanley Institute for Cognitive Genomics

Acknowledgements

Cold Spring Harbor Laboratory
Stanley Institute for Cognitive Genomics

Jason O'Rawe



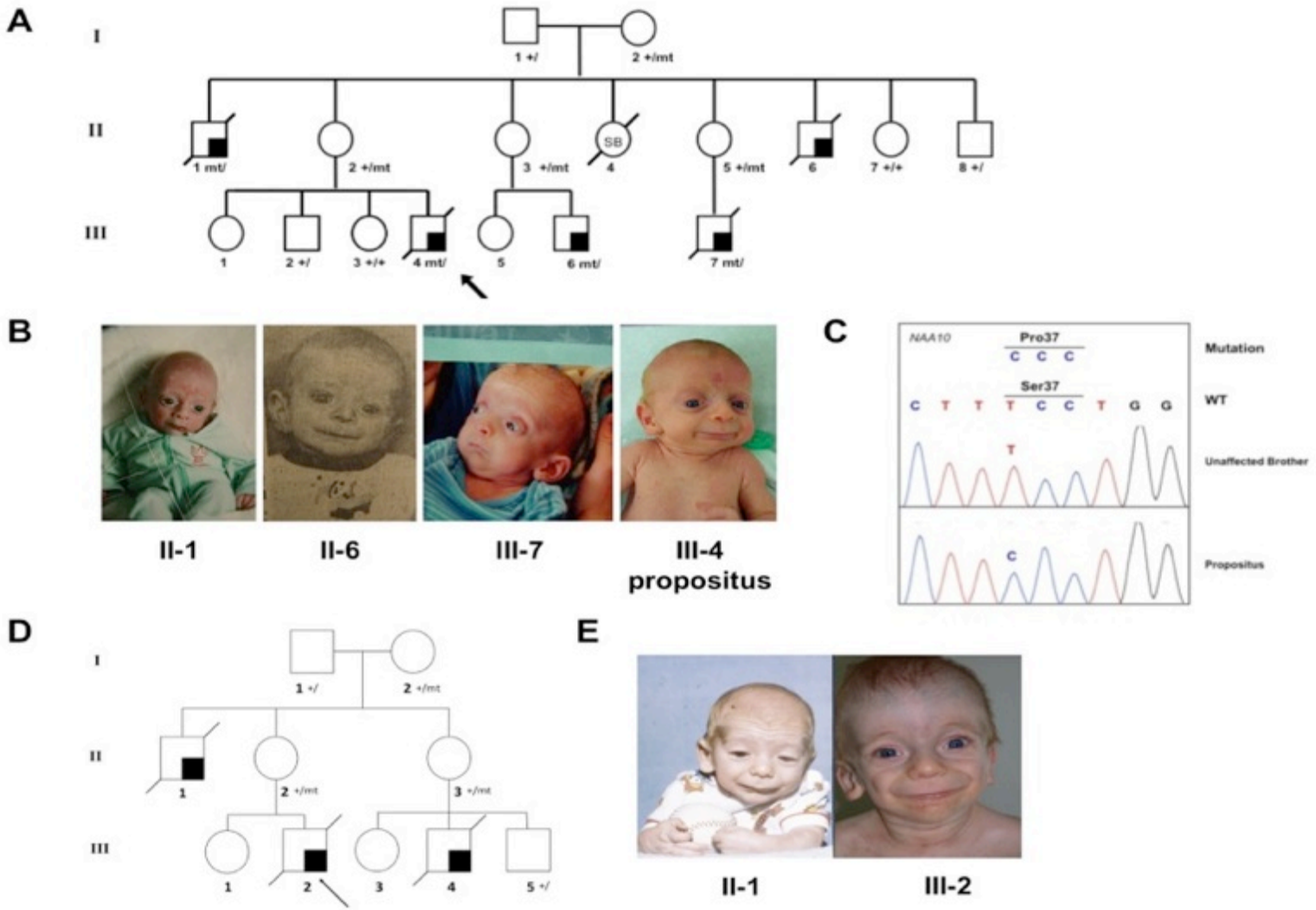
Yiyang Wu

Kai Wang

Michael Schatz
Giuseppe Narzisi
Eric Antoniou
Dick McCombie
Sequencing core facility

Tao Jiang
Guangqing Sun

Ogden Syndrome – in 2011



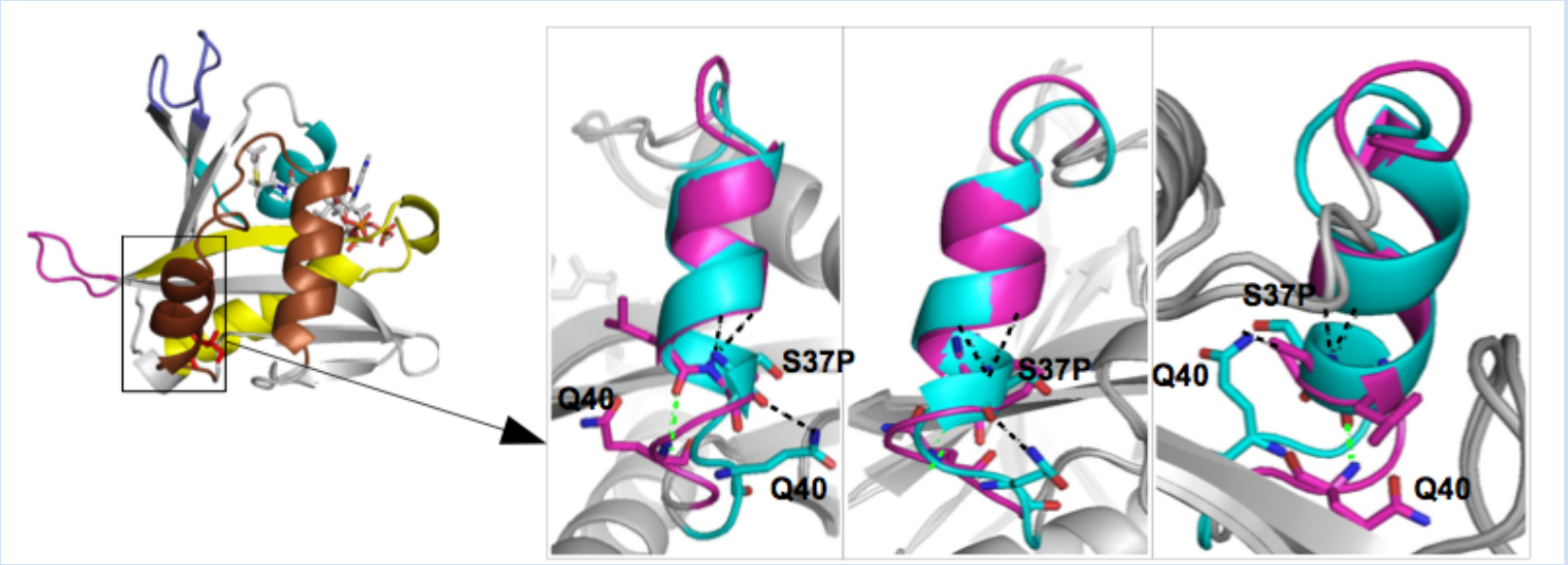
We found the SAME mutation in two unrelated families, with a very similar phenotype in both families, helping prove that this genotype contributes to the phenotype observed.

These are the Major Features of the Syndrome.

Table 1. Features of the syndrome	
Growth	post-natal growth failure
Development	global, severe delays
Facial	prominence of eyes, down-sloping palpebral fissures, thickened lids large ears beaking of nose, flared nares, hypoplastic alae, short columella protruding upper lip micro-retrognathia
Skeletal	delayed closure of fontanels broad great toes
Integument	redundancy / laxity of skin minimal subcutaneous fat cutaneous capillary malformations
Cardiac	structural anomalies (ventricular septal defect, atrial level defect, pulmonary artery stenoses) arrhythmias (Torsade de points, PVCs, PACs, SVtach, Vtach) death usually associated with cardiogenic shock preceded by arrythmia.
Genital	inguinal hernia hypo- or cryptorchidism
Neurologic	hypotonia progressing to hypertonia cerebral atrophy neurogenic scoliosis
Shaded regions include features of the syndrome demonstrating variability. Though variable findings of the cardiac, genital and neurologic systems were observed, all affected individuals manifested some pathologic finding of each.	

The mutation is a missense resulting in Serine to Proline change in Naa10p

- Ser 37 is conserved from yeast to human
- Ser37Pro is predicted to affect functionality (SIFT and other prediction programs)
- Structural modelling of hNaa10p wt (cyan) and S37P (pink)

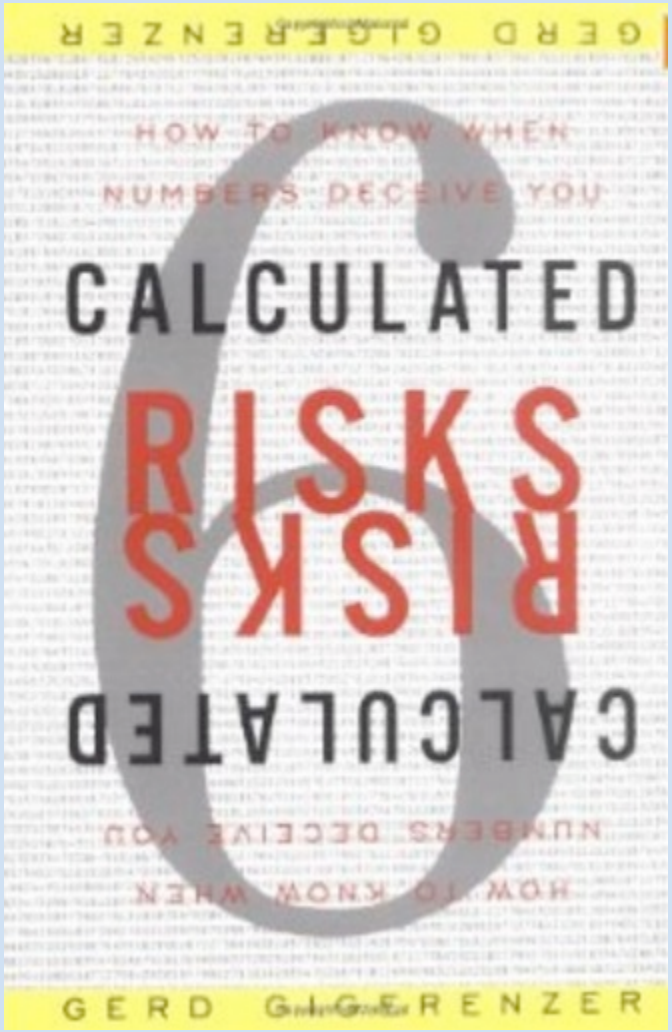


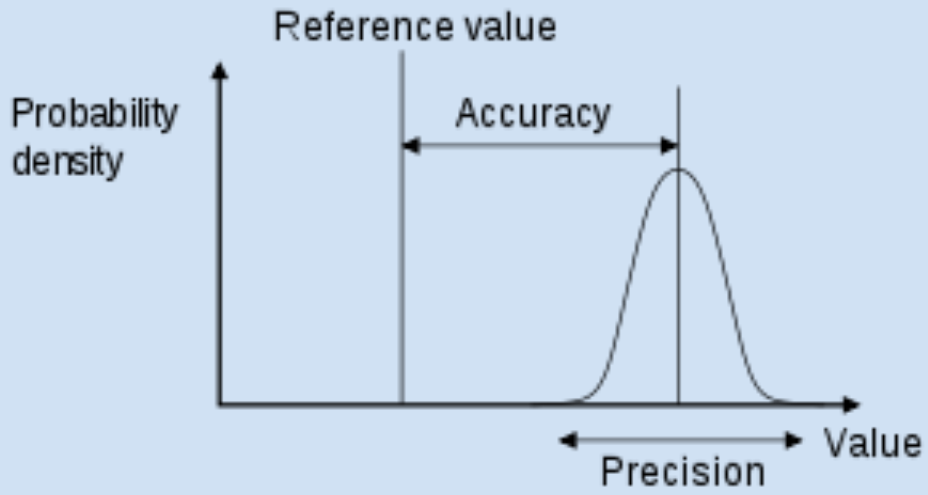
Results from Exome and WGS requires both Analytic and Clinical Validity

- Analytical Validity: the test is accurate with high sensitivity and specificity.

- Clinical Validity: Given an accurate test result, what impact and/or outcome does this have on the individual person?

Illusions of Certainty. Everything is Probabilistic.


GERD GIGERENZER
HOW TO KNOW WHEN
NUMBERS DECEIVE YOU
CALCULATED
RISKS
RISKS
CALCULATED
NUMBERS DECEIVE YOU
HOW TO KNOW WHEN
GERD GIGERENZER



High accuracy, but low precision

High precision, but low accuracy

In the fields of science, engineering, industry, and statistics, the **accuracy** of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The **precision** of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

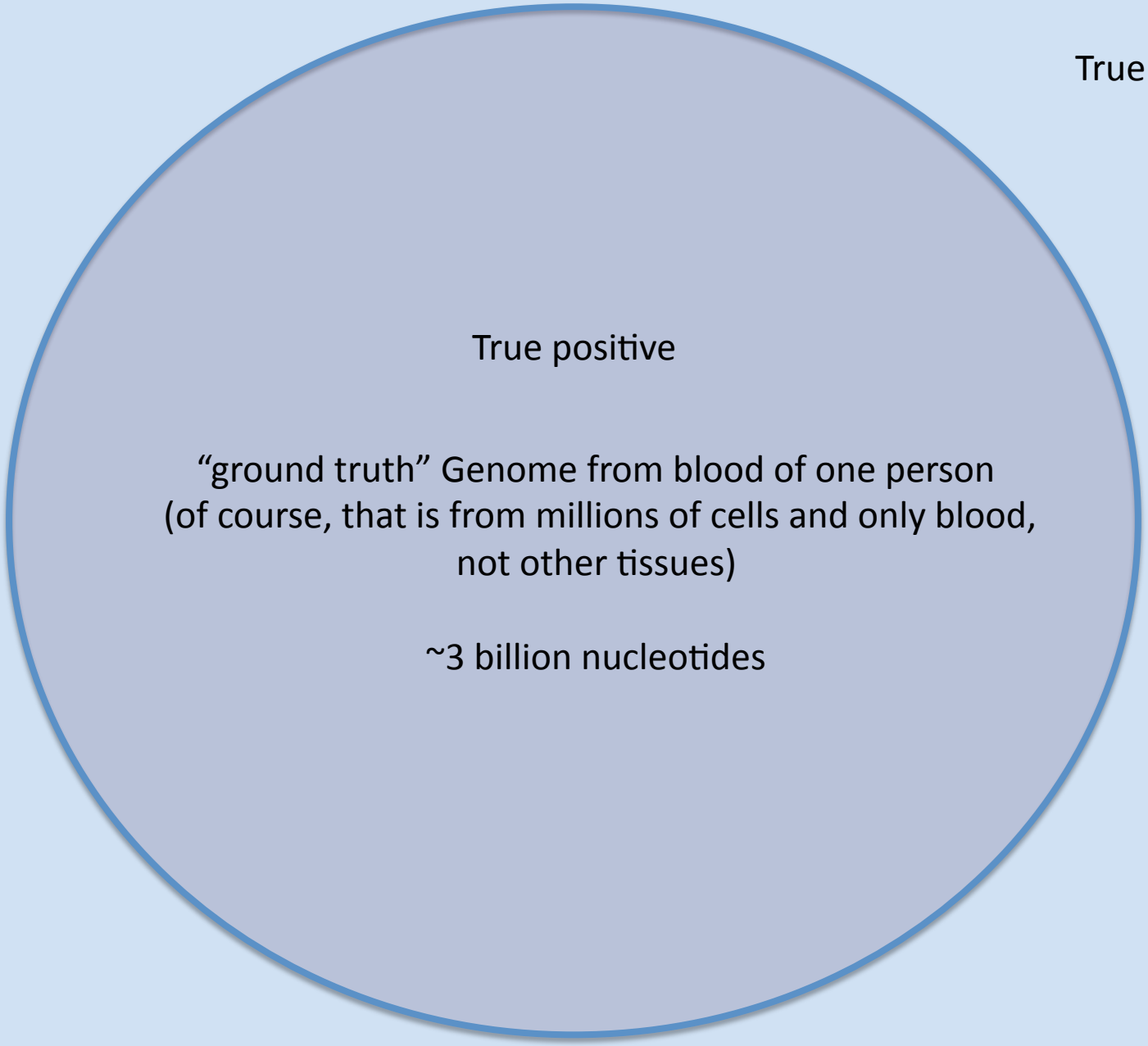
http://en.wikipedia.org/wiki/Accuracy_and_precision

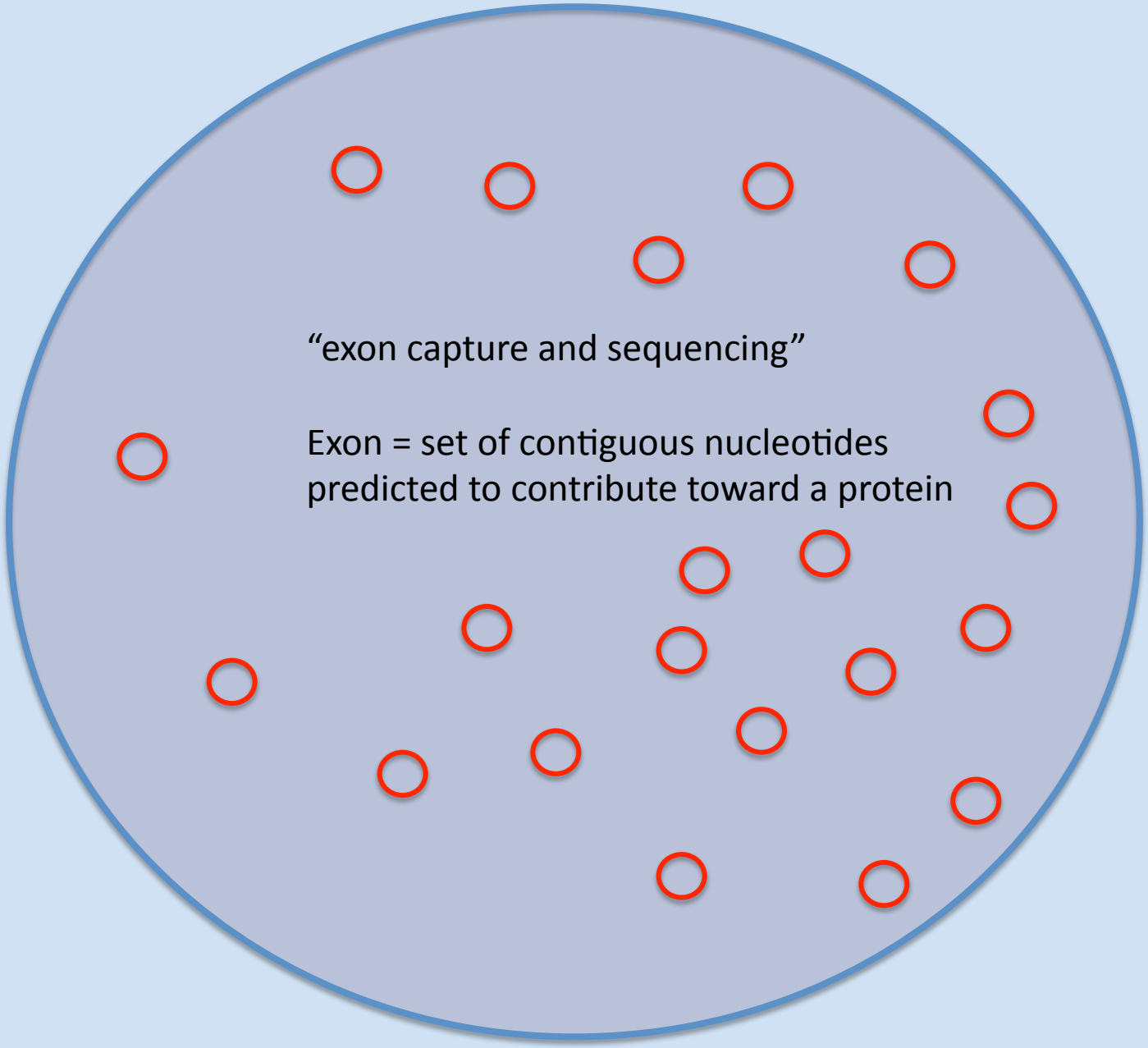
	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)		

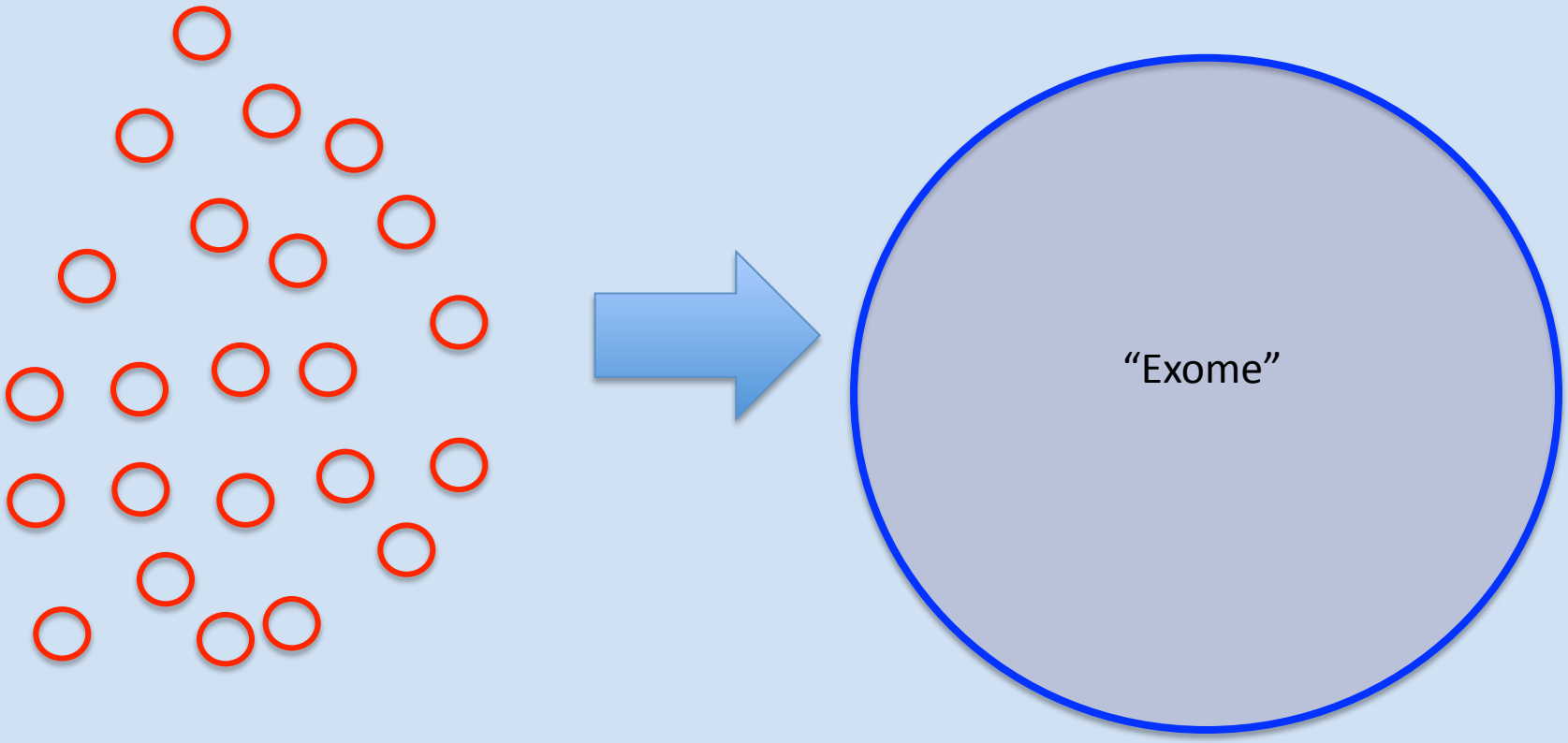
Accuracy

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

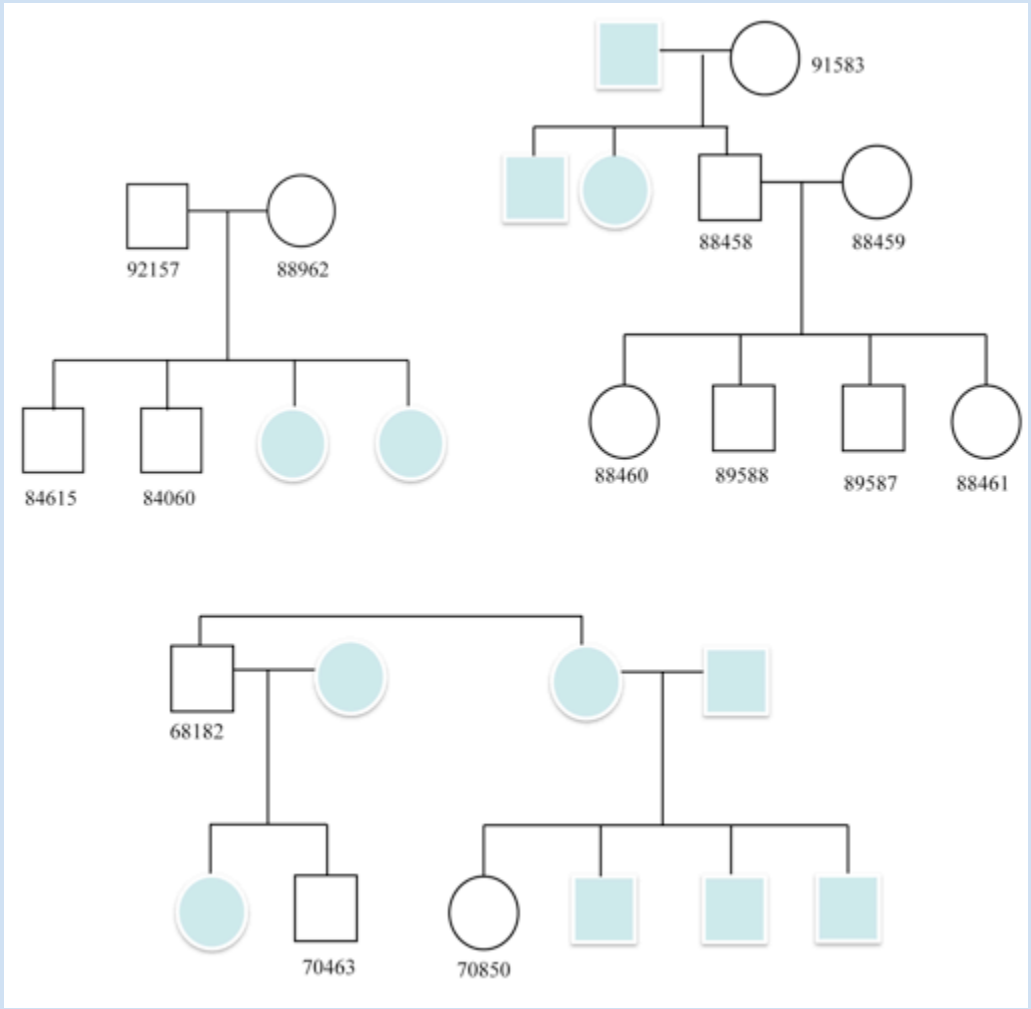
An accuracy of 100% means that the measured values are exactly the same as the given values.


True negative
True positive
“ground truth” Genome from blood of one person
(of course, that is from millions of cells and only blood,
not other tissues)
~3 billion nucleotides


“exon capture and sequencing”
Exon = set of contiguous nucleotides
predicted to contribute toward a protein


“Exome”

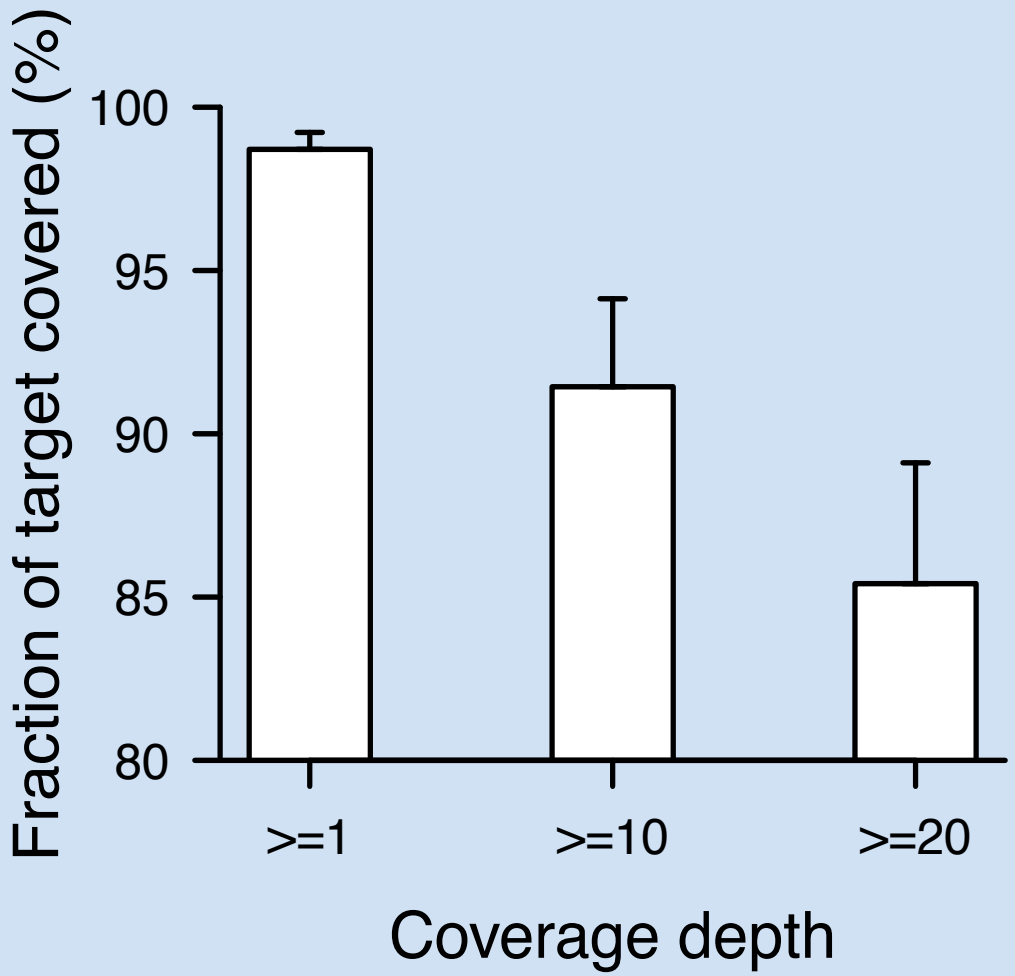
Chose to sequence 15 “exomes”



2-3 rounds of sequencing at BGI to attain goal of >80% of target region at >20 reads per base pair

Exome Capture Statistics	K24510-84060	K24510-92157-a	K24510-84615	K24510-88962
Target region (bp)	46,401,121	46,401,121	46,401,121	46,257,379
Raw reads	138,779,950	161,898,170	156,985,870	104,423,704
Raw data yield (Mb)	12,490	14,571	14,129	9,398
Reads mapped to genome	110,160,277	135,603,094	135,087,576	83,942,646
Reads mapped to target region	68,042,793	84,379,239	80,347,146	61,207,116
Data mapped to target region (Mb)	5,337.69	6,647.18	6,280.01	4,614.47
Mean depth of target region	**115.03**	**143.25**	**135.34**	**99.76**
Coverage of target region (%)	**0.9948**	**0.9947**	**0.9954**	**0.9828**
Average read length (bp)	89.91	89.92	89.95	89.75
Fraction of target covered >=4X	98.17	98.38	98.47	94.25
Fraction of target covered >=10X	95.18	95.90	95.97	87.90
Fraction of target covered >=20X	**90.12**	**91.62**	**91.75**	**80.70**
Fraction of target covered >=30X	84.98	87.42	87.67	74.69
Capture specificity (%)	61.52	62.12	59.25	73.16
Fraction of unique mapped bases on or near target	65.59	65.98	63.69	85.46
Gender test result	M	M	M	F

Depth of Coverage in 15 exomes > 20 reads per bp in target region



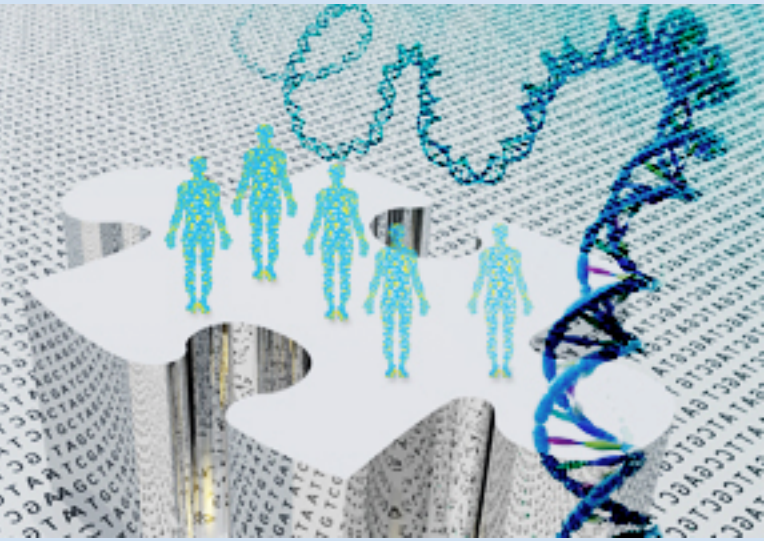
Experimental Design

- Evaluate robustness of variant calling implemented by different bioinformatics analysts.
- Looking at False Positives and False Negatives.
- How reliable are variants that are uniquely called by individual pipelines?
- Are some pipelines better at detecting rare, or novel variants than others?

Human Exome Sequencing Promotion

50X: $899/sample 100X: $1299/sample (SNP & Indel Included)

38,000 Exomes Sequenced by BGI to Date



Promotion Details (valid for Americas and Europe customers NOW through MAY 31)

A. The 899 USD/sample package – 50X human exome sequencing

Agilent SureSelect 50/51M Capture kit
100 bp paired-end sequencing on HiSeq 2000
5 Gb high quality* sequencing data
50X average coverage for target regions guaranteed
SNP & Indel calling and annotation included

B. The 1299 USD/sample package – 100X human exome sequencing

Agilent SureSelect 50/51M Capture kit
100 bp paired-end sequencing on HiSeq 2000
10 Gb high quality* sequencing data
100X average coverage for target regions guaranteed
SNP & Indel calling and annotation included

Table 1. A descriptive summary of the variant calling pipelines included in the comparative analyses.

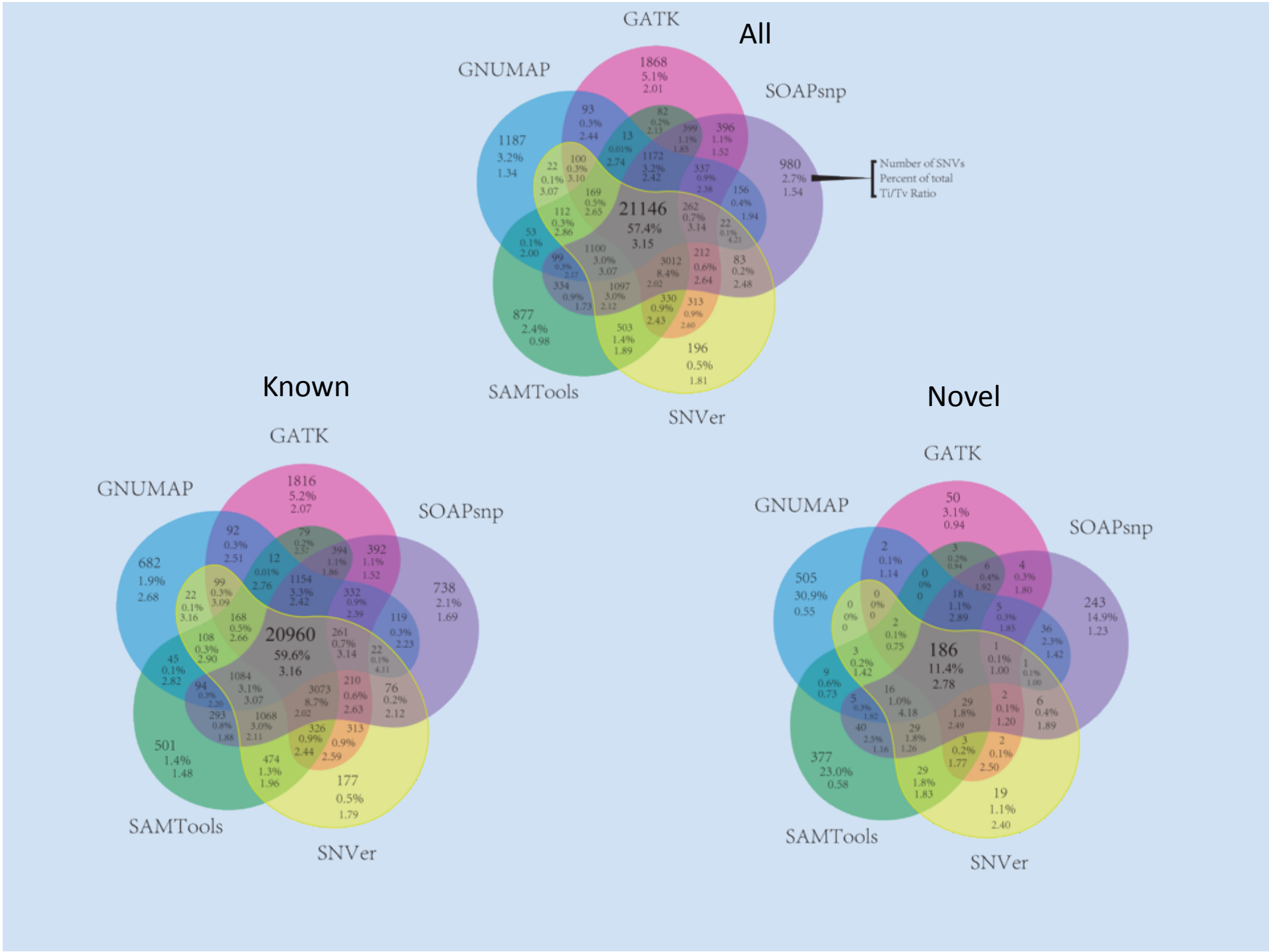
Pipeline name	Alignment method	Variant-calling module	Description of variant-calling algorithm
SOAP	SOAPaligner version 2.21/ BWA version 0.5.9	SOAPsnp version 1.03/ SOAPindel version 2.01	SOAP uses a method based on Bayes' theorem to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors [22].
GATK version 1.5	BWA version 0.5.9	UnifiedGenotyper version 1.5	GATK employs a general Bayesian framework to distinguish and call variants. Error correction models are guided by expected characteristics of human variation to further refine variant calls [19].
SNVer version 0.2.1	BWA version 0.5.9	SNVer version 0.2.1	SNVer uses a more general frequentist framework, and formulates variant calling as a hypothesis-testing problem [25].
GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP incorporates the base uncertainty of the reads into mapping analysis using a probabilistic Needleman-Wunsch algorithm [24].
SAMtools version 0.1.18	BWA version 0.5.9	mpileup version 0.1.18	SAMtools [20] calls variants by generating a consensus sequence using the MAQ model framework, which uses a general Bayesian framework for picking the base that maximizes the posterior probability with the highest Phred quality score.

Accuracy ⟷ Precision

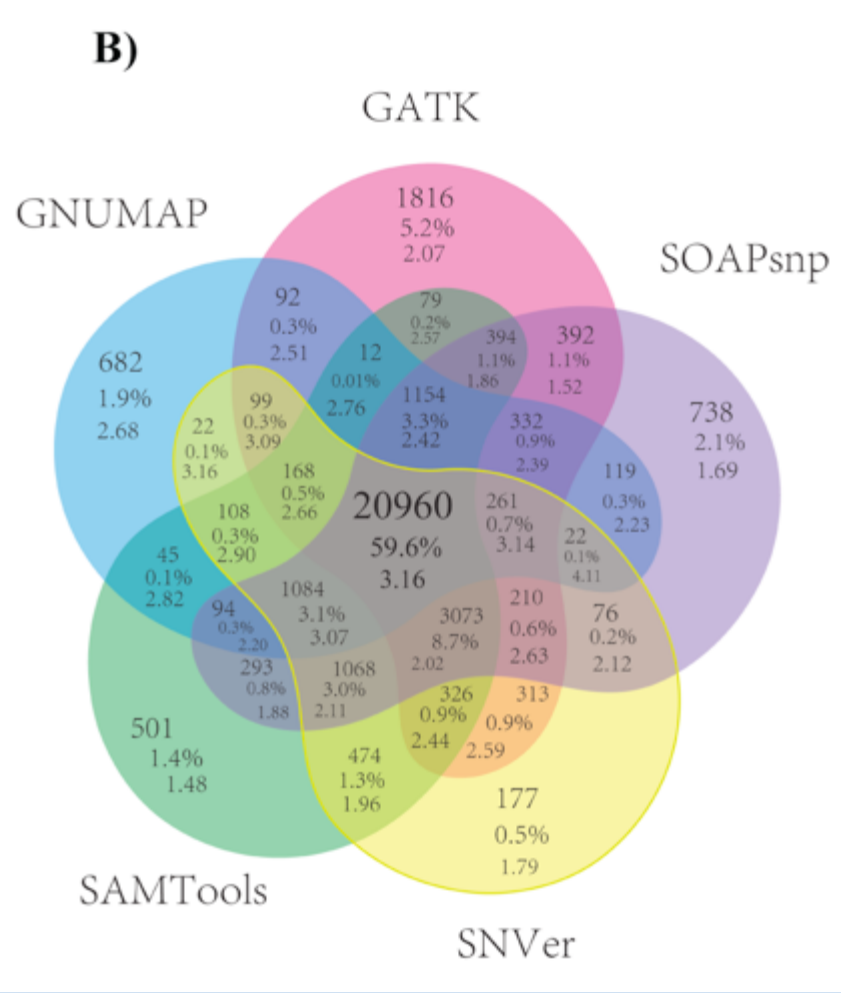
Error-Prone & Doubtful

Repeatable & Reproducible

Images: Wikipedia

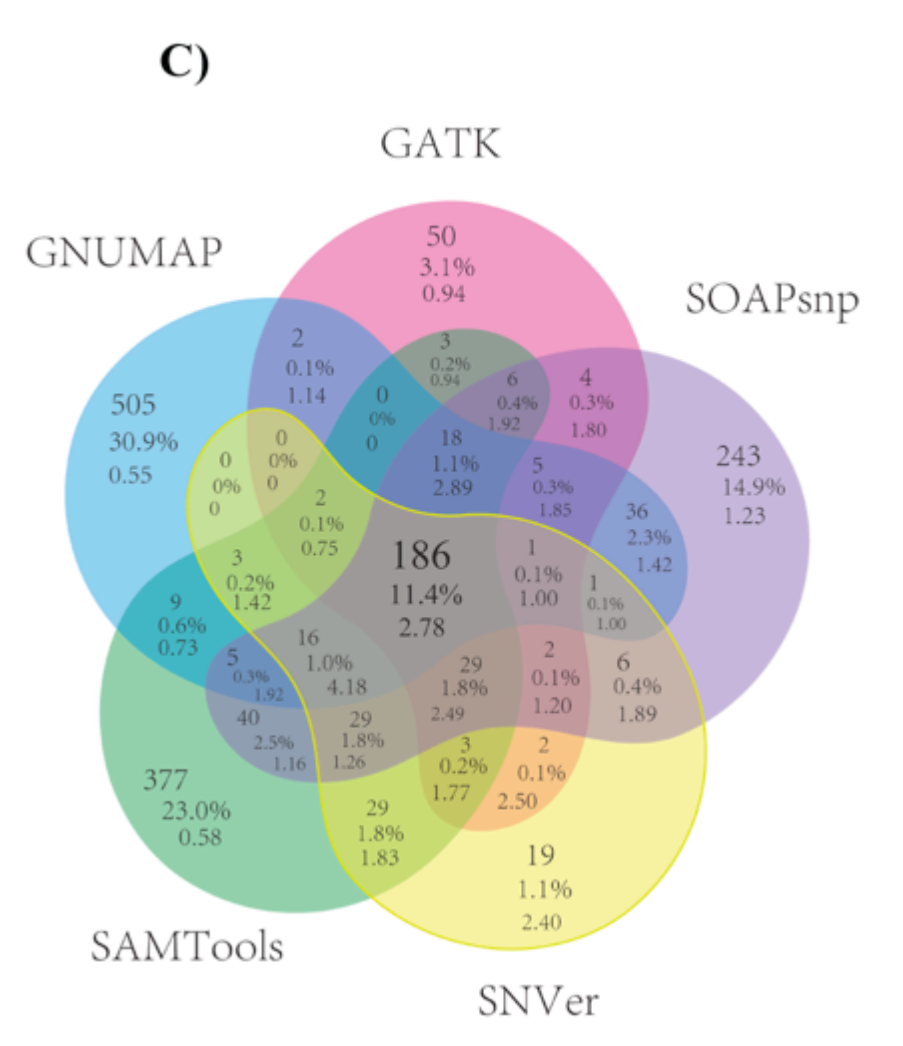

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
877 2.4% 0.98
196 0.5% 1.81
21146 57.4% 3.15
Number of SNVs
Percent of total
Ti/Tv Ratio
Known
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1816 5.2% 2.07
682 1.9% 2.68
738 2.1% 1.69
501 1.4% 1.48
177 0.5% 1.79
20960 59.6% 3.16
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50 3.1% 0.94
505 30.9% 0.55
243 14.9% 1.23
377 23.0% 0.58
19 1.1% 2.40
186 11.4% 2.78

Known SNVs



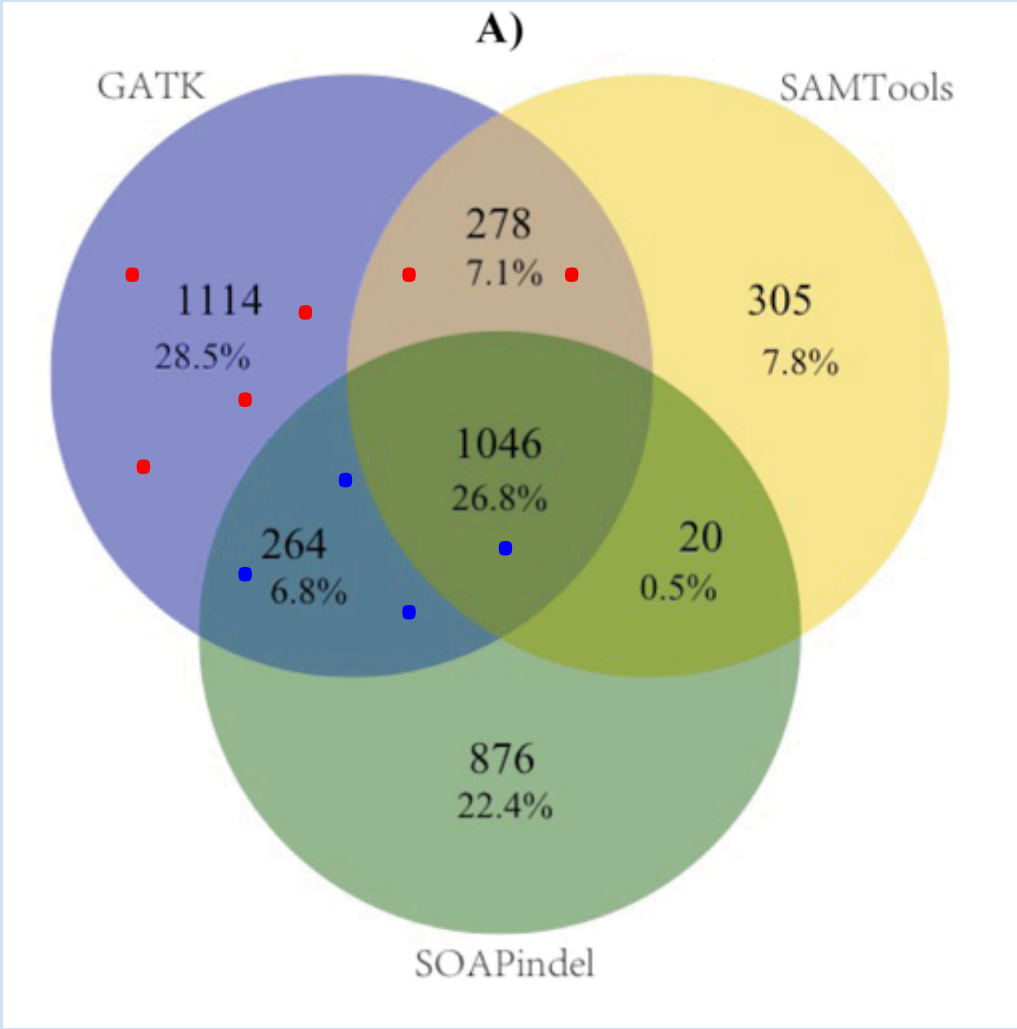
B) Mean # of known SNVs (present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the the Venn diagram is the percent of known SNVs called by all five pipelines.

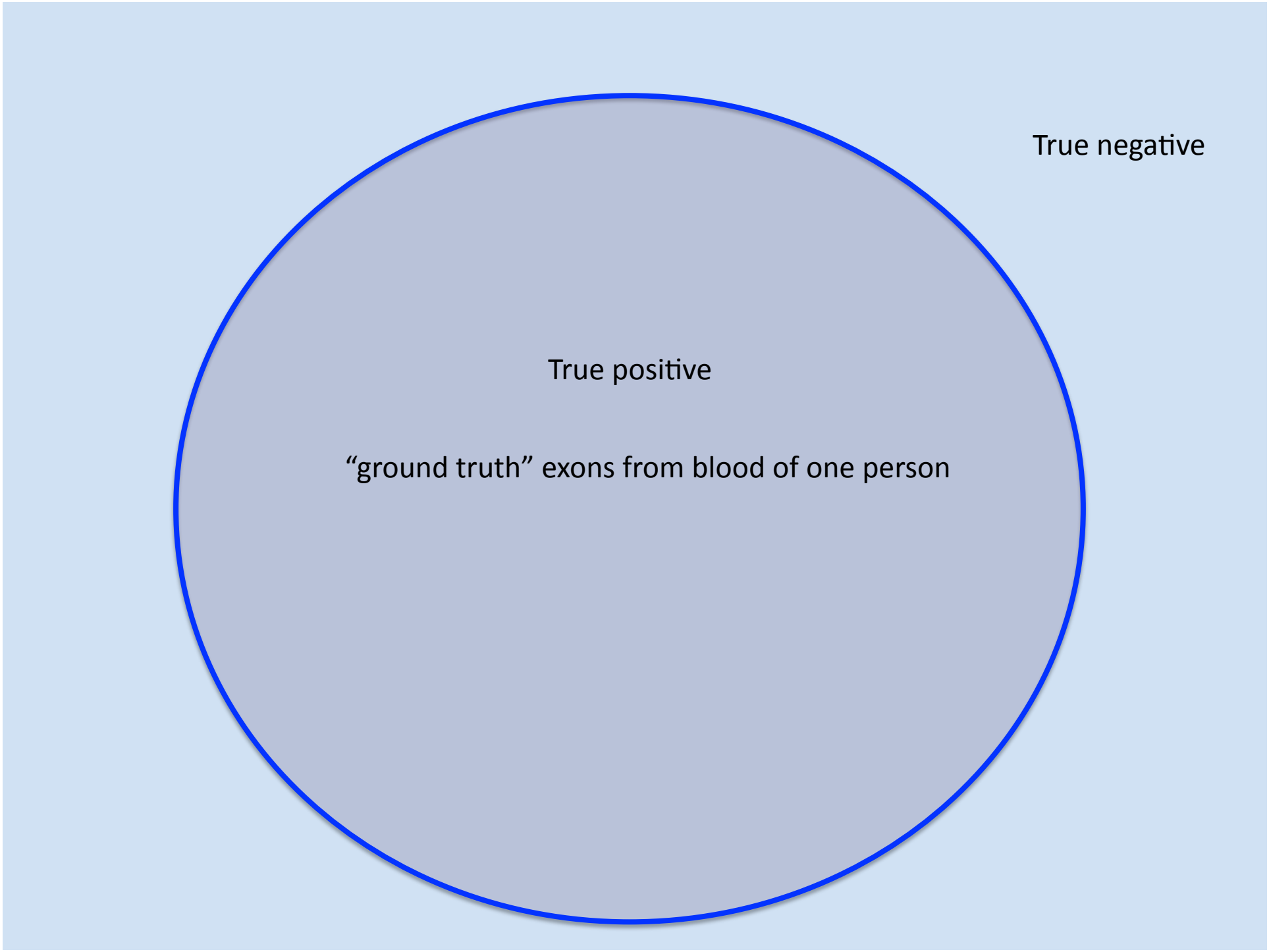
Novel SNVs

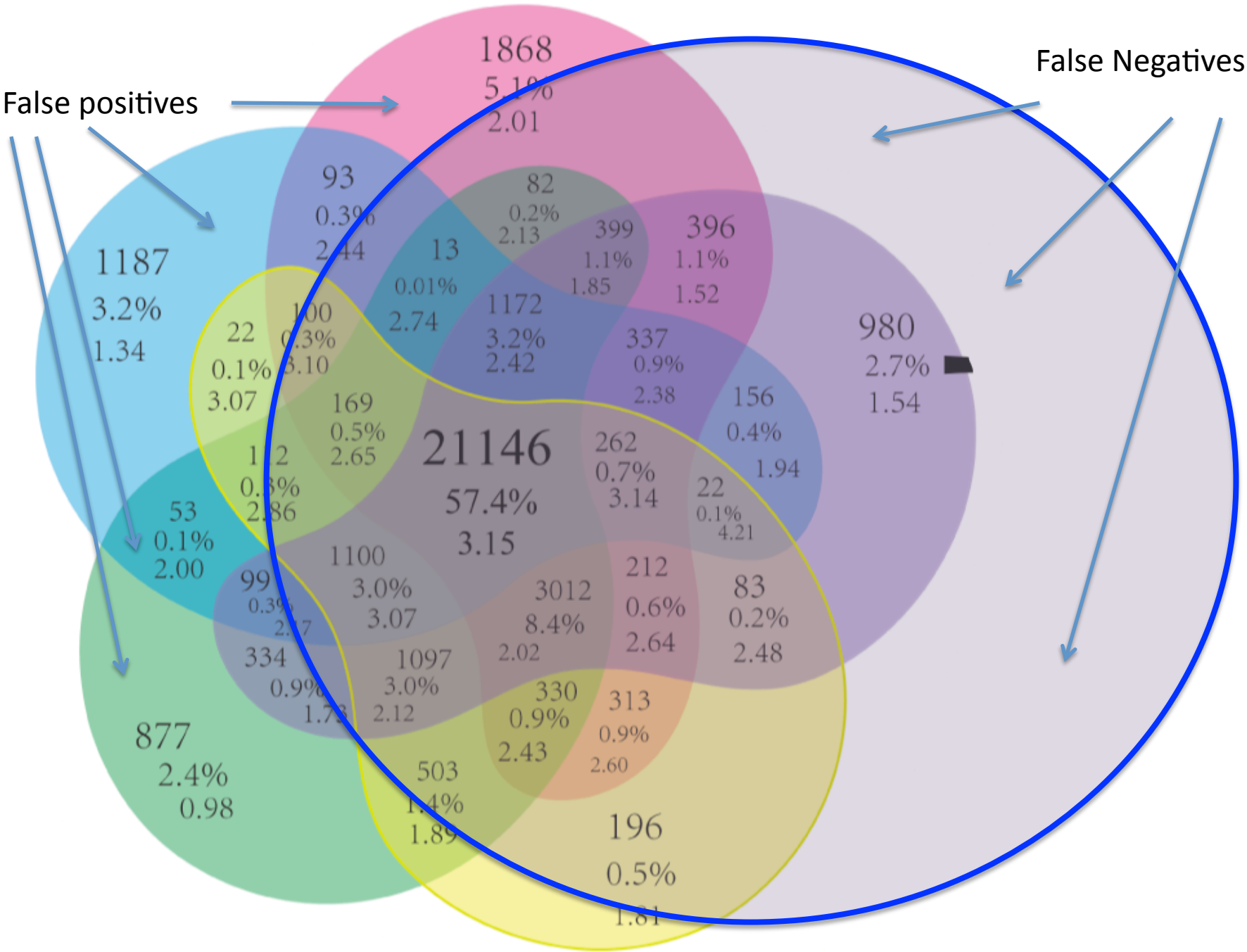


C) Mean # of novel SNVs (not present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the Venn diagram is the percent of novel SNVs called by all five pipelines.

Indels called by GATK, SOAP and SAMtools

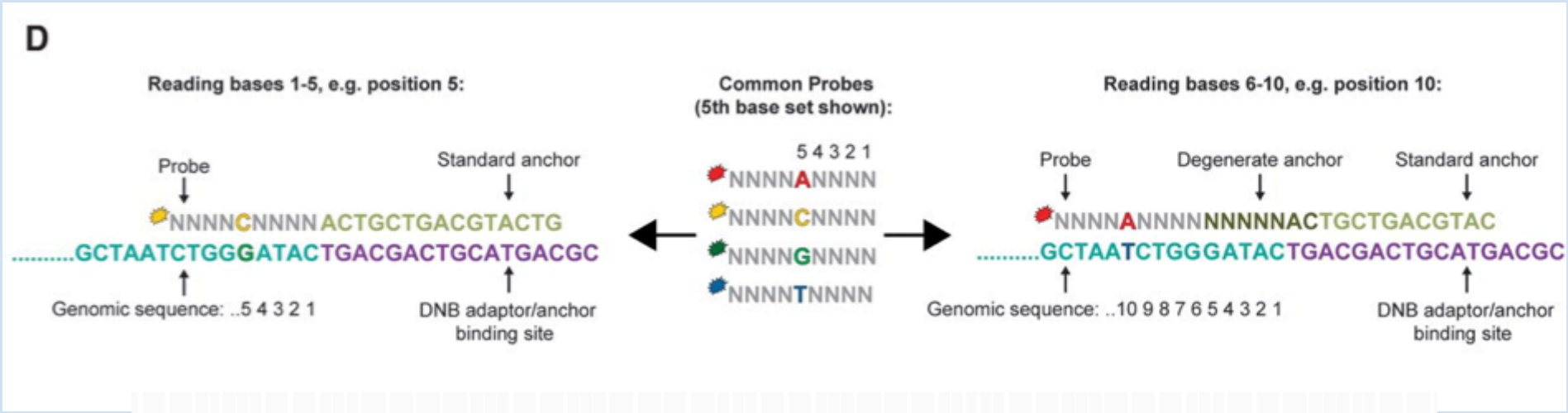


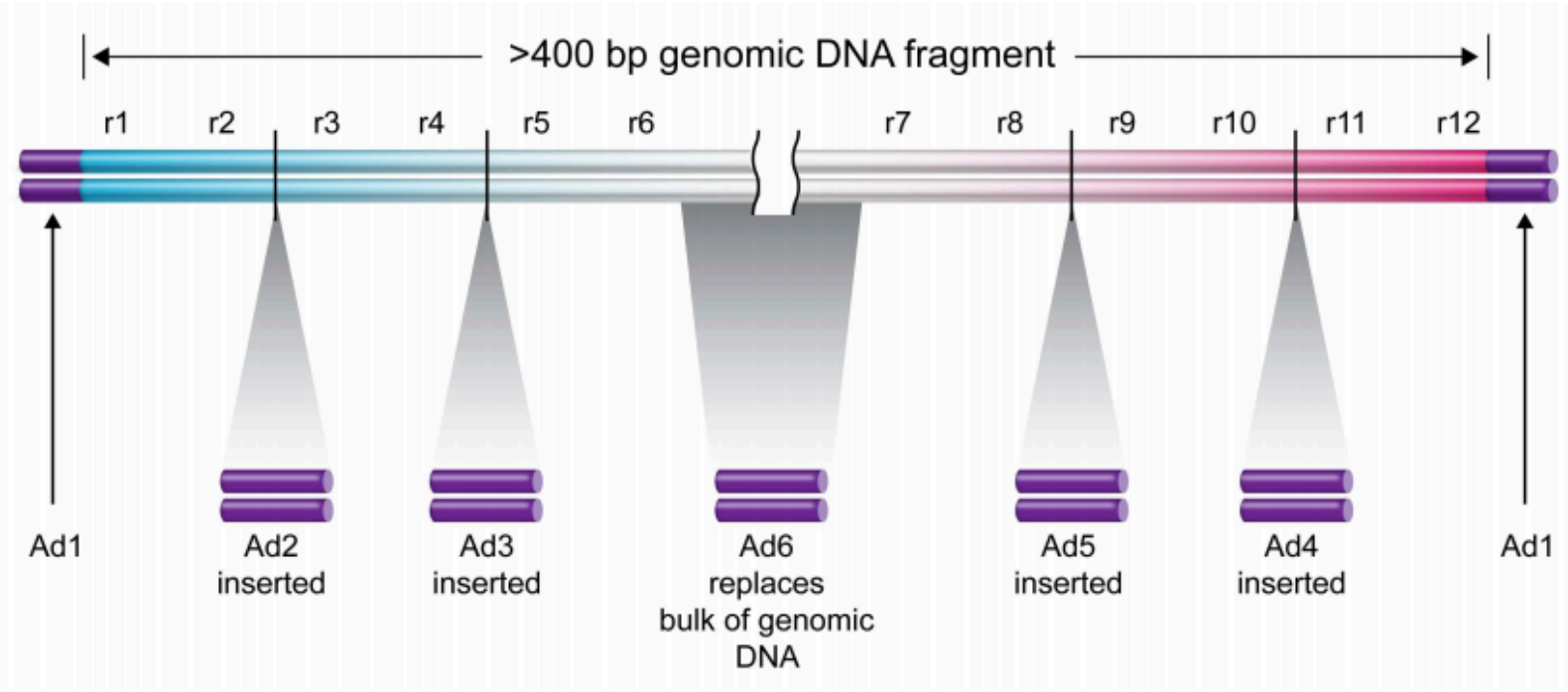

True negative
True positive
“ground truth” exons from blood of one person


False positives
False Negatives
1868
5.1%
2.01
93
0.3%
2.44
82
0.2%
2.13
399
1.1%
1.85
396
1.1%
1.52
1187
3.2%
1.34
13
0.01%
2.74
1172
3.2%
2.42
337
0.9%
2.38
980
2.7%
1.54
22
0.1%
3.07
100
0.3%
3.10
169
0.5%
2.65
156
0.4%
1.94
21146
57.4%
3.15
262
0.7%
3.14
22
0.1%
4.21
53
0.1%
2.00
1100
3.0%
3.07
212
0.6%
2.64
83
0.2%
2.48
3012
8.4%
2.02
334
0.9%
1097
3.0%
2.12
330
0.9%
2.43
313
0.9%
2.60
877
2.4%
0.98
503
1.4%
196
0.5%
1.81

Cross validation using orthogonal sequencing technology (Complete Genomics)

Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)





Accuracy of Complete Genomics Whole Human Genome Sequencing Data

Analysis Pipeline v2.0

	FALSE POSITIVES	EST FPs	FALSE NEGATIVES	TOTAL DISCORDANCES	CONCORDANCE
Discordant SNVs per called MB	1.56 x 10-6	4,450	1.67 x 10-6	3.23 x 10-6	99.9997% of bases

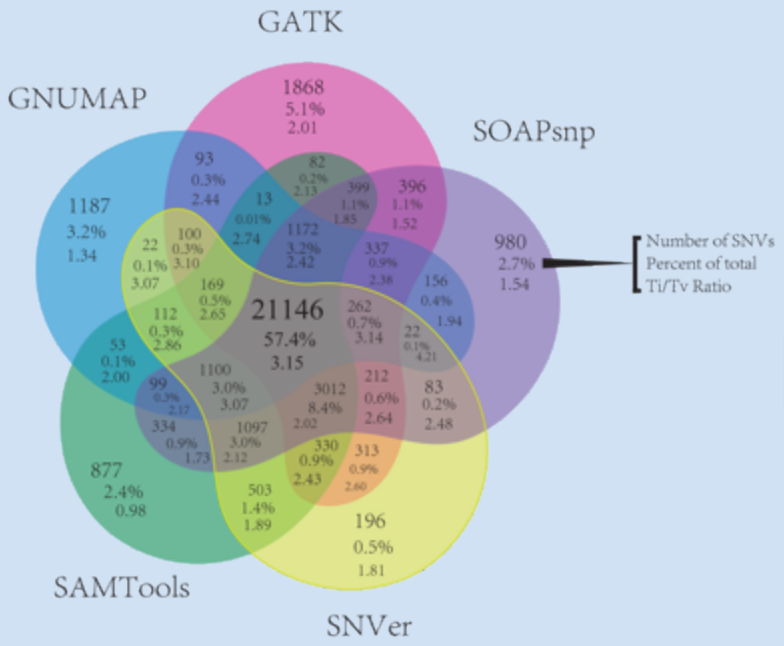
Table 2. *Concordance of Technical Replicates.*

COMPLETE GENOMICS CALL	OTHER PLATFORM	PLATFORM-SPECIFIC SNVs	VALIDATION RATE	EST FPs	FPR
Het or Hom SNV	No SNV Reported	99K	17/18 = 94.4%	5,577	0.16%
No-call or Hom-Ref	SNV Reported	345K	2/15 = 13.3%	299,115	8.2%

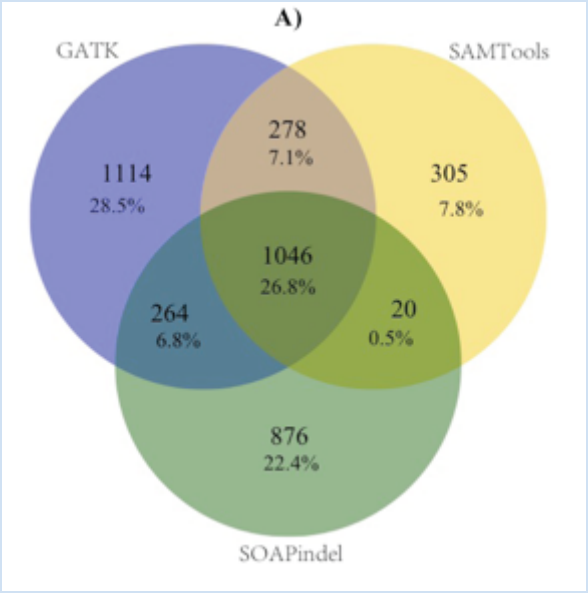
Table 3. *False Positive Rate.*

Performance comparison of whole-genome sequencing platforms

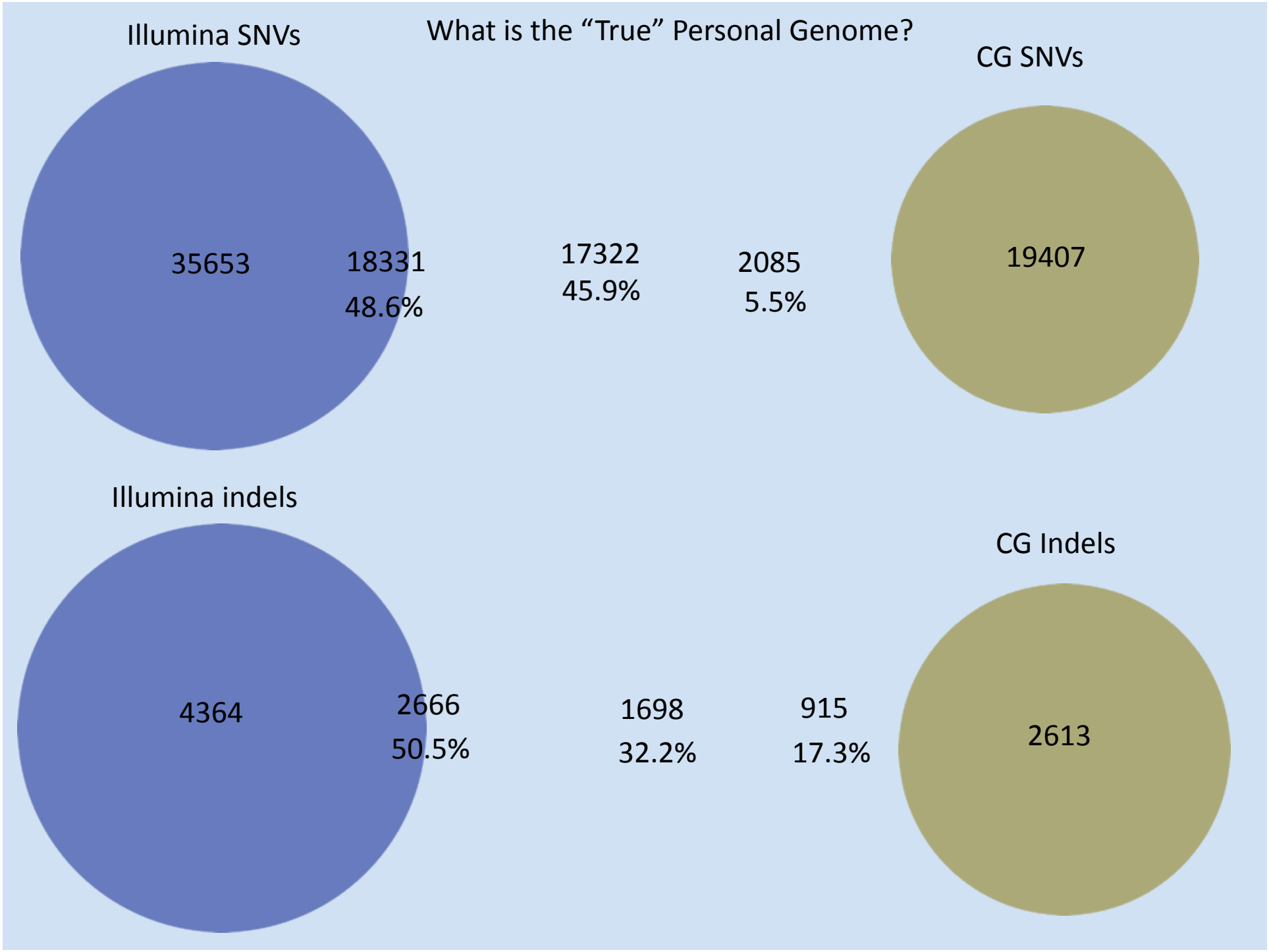
Hugo Y K Lam[1,8], Michael J Clark[1], Rui Chen[1], Rong Chen[2,8], Georges Natsoulis[3], Maeve O'Huallachain[1], Frederick E Dewey[4], Lukas Habegger[5], Euan A Ashley[4], Mark B Gerstein[5–7], Atul J Butte[2], Hanlee P Ji[3] & Michael Snyder[1]

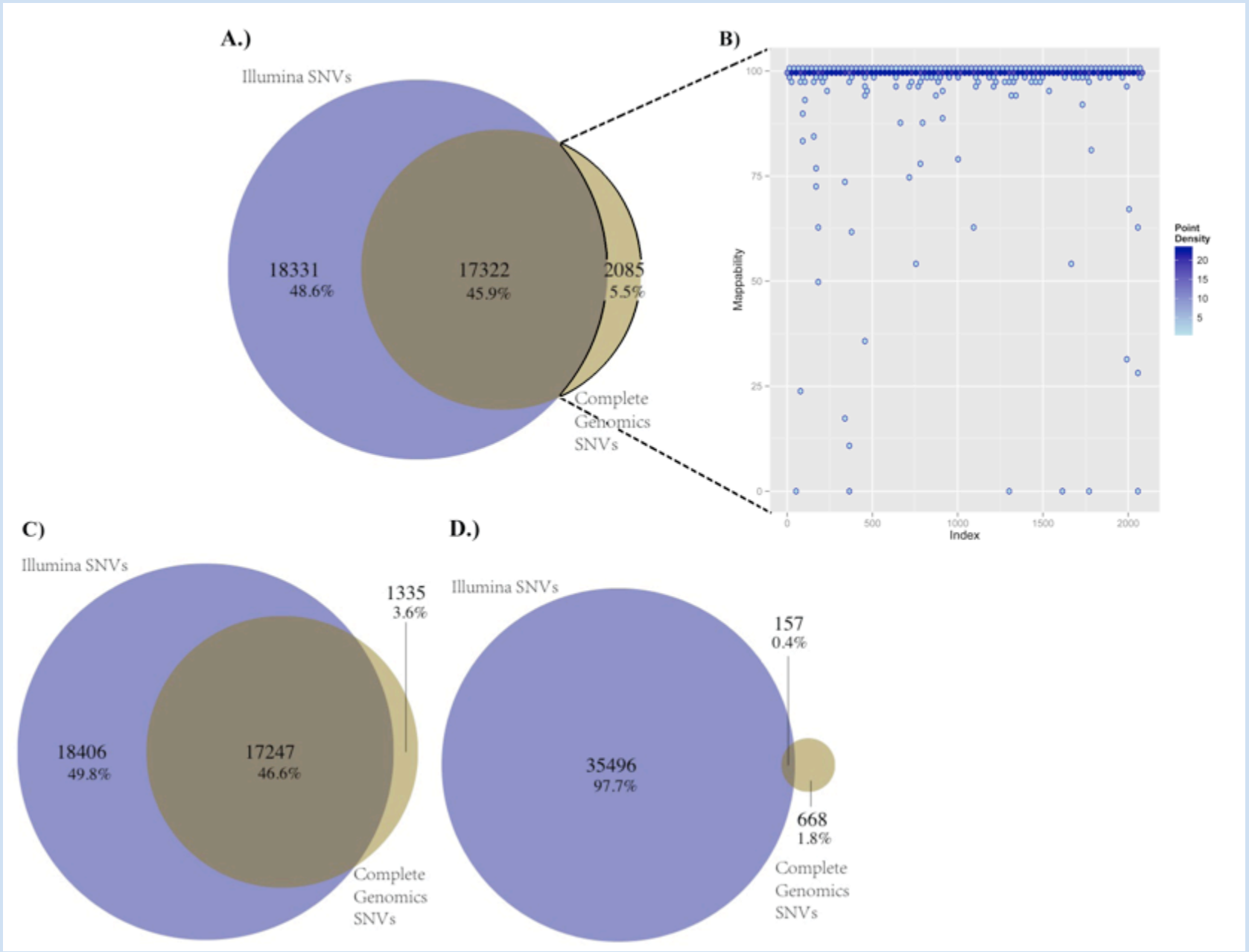


Union of all SNVs from 5 pipelines



Union of all indels from 3 pipelines


What is the “True” Personal Genome?
Illumina SNVs
CG SNVs
35653
18331
48.6%
17322
45.9%
2085
5.5%
19407
Illumina indels
CG Indels
4364
2666
50.5%
1698
32.2%
915
17.3%
2613


A.)
Illumina SNVs
18331
48.6%
17322
45.9%
2085
5.5%
Complete
Genomics
SNVs
B)
Mappability
Index
Point
Density
20
15
10
5
C)
Illumina SNVs
18406
49.8%
17247
46.6%
1335
3.6%
Complete
Genomics
SNVs
D.)
Illumina SNVs
35496
97.7%
157
0.4%
668
1.8%
Complete
Genomics
SNVs

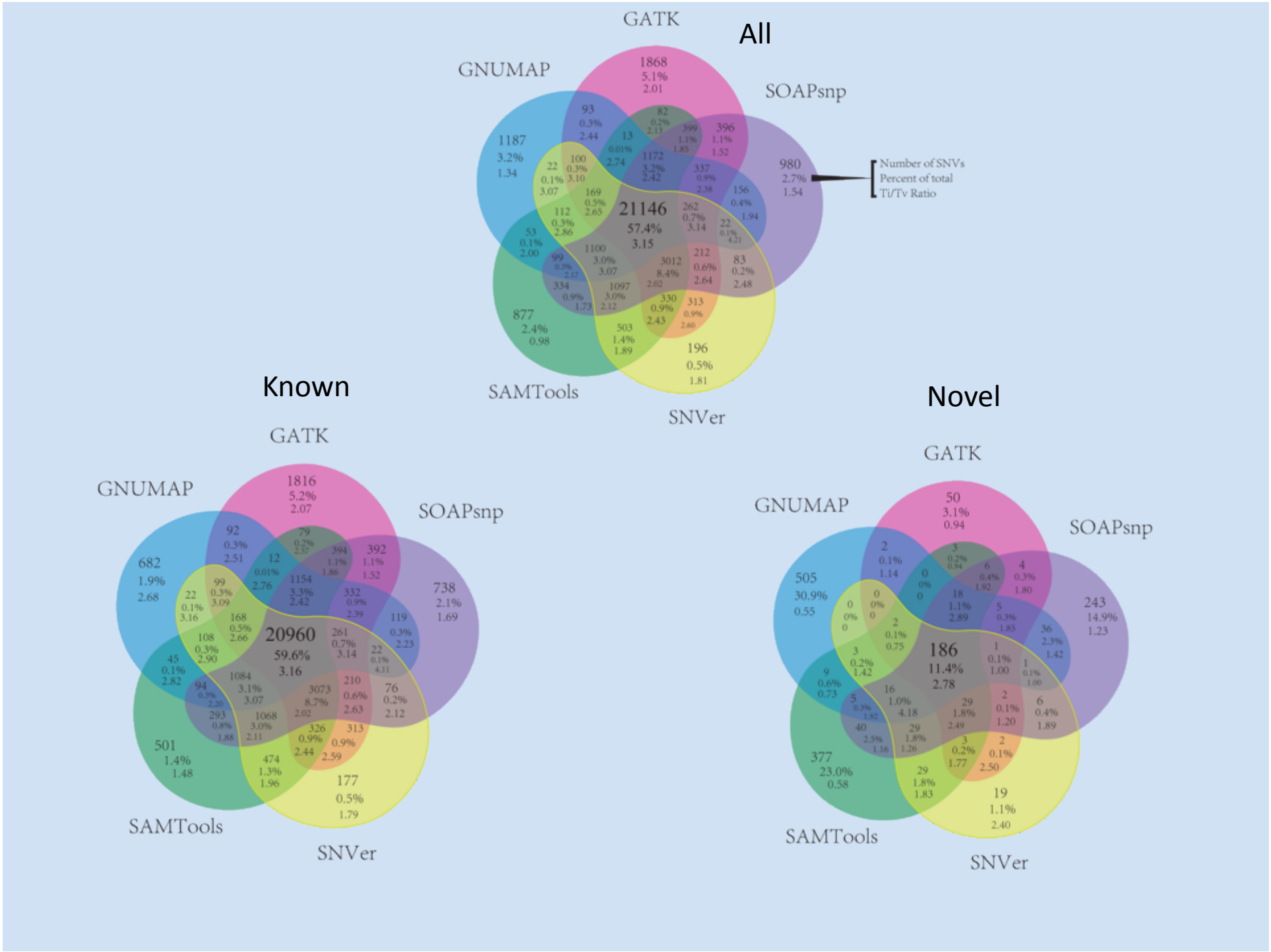
Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score

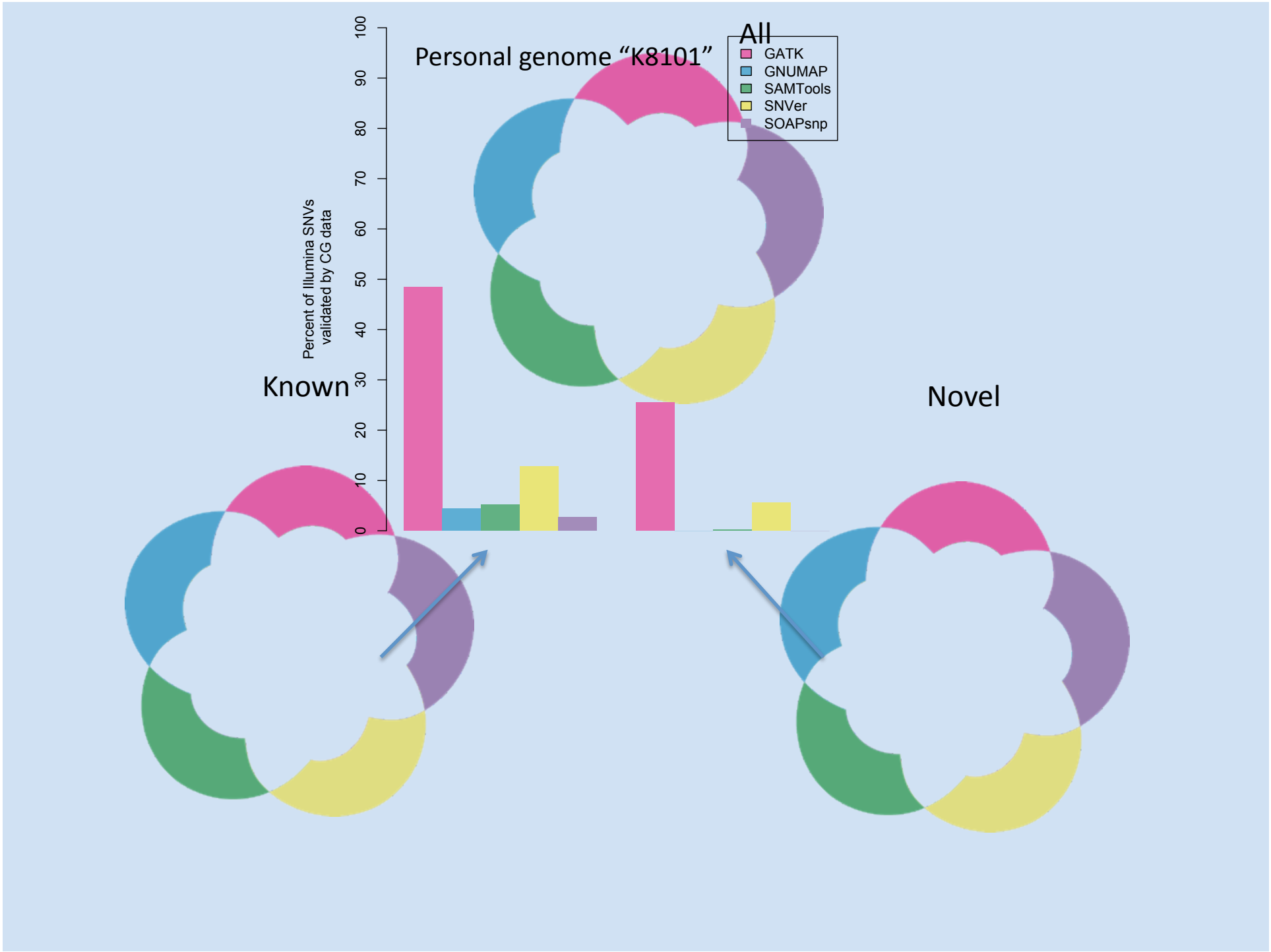
Hayan Lee[1,2*] and Michael C. Schatz [1,2]

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY
[2]Simons Center for Quantitive Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Bioinformatics Advance Access published June 4, 2012

- Genome Mappability Score (GMS) -- measure of the complexity of resequencing a genome = a weighted probability that any read could be unambiguously mapped to a given position, and thus measures the overall composition of the genome itself.

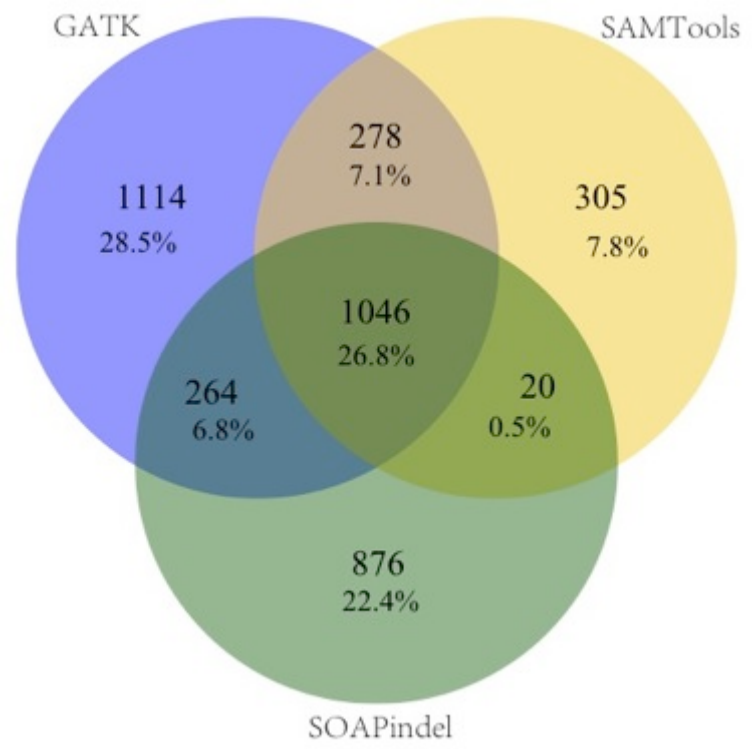

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
877 2.4% 0.98
196 0.5% 1.81
21146 57.4% 3.15
Number of SNVs
Percent of total
Ti/Tv Ratio
Known
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1816 5.2% 2.07
682 1.9% 2.68
738 2.1% 1.69
501 1.4% 1.48
177 0.5% 1.79
20960 59.6% 3.16
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50 3.1% 0.94
505 30.9% 0.55
243 14.9% 1.23
377 23.0% 0.58
19 1.1% 2.40
186 11.4% 2.78

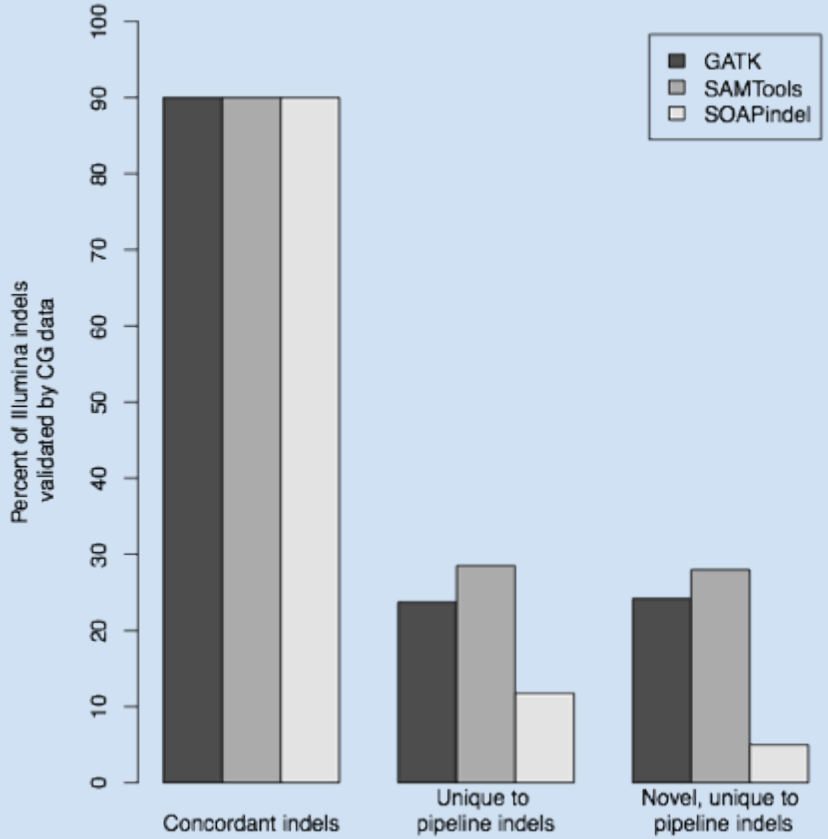

Personal genome "K8101"
All
GATK
GNUMAP
SAMTools
SNVer
SOAPsnp
Percent of Illumina SNVs validated by CG data
0
10
20
30
40
50
60
70
80
90
100
Known
Novel

Higher Validation by CG of SNVs with the BWA-GATK(v1.5) pipeline

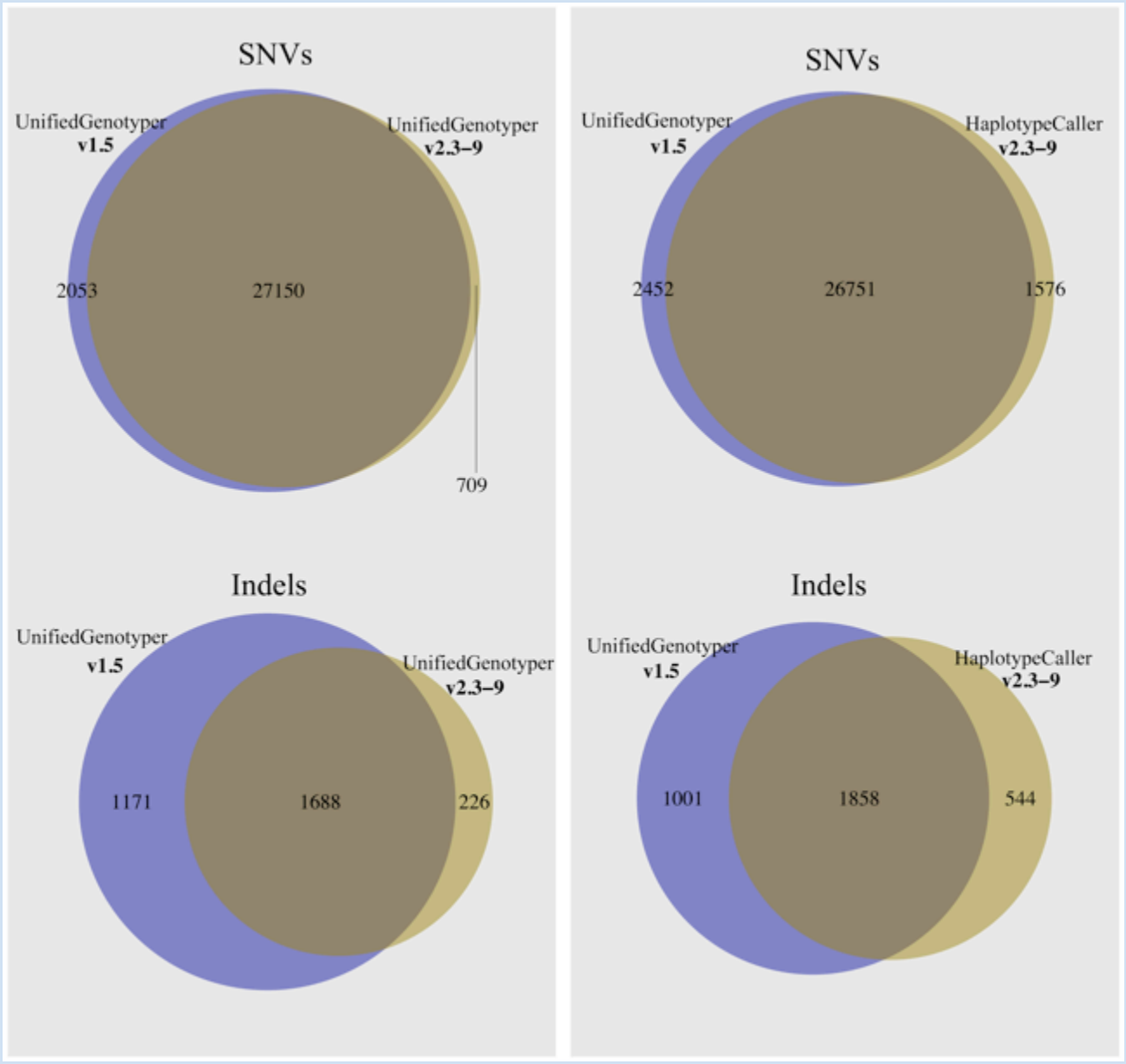
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK(v1.5), in comparison to the other 4 pipelines (including SOAP).

Validating Indels with Complete Genomics Data for the 3 pipelines

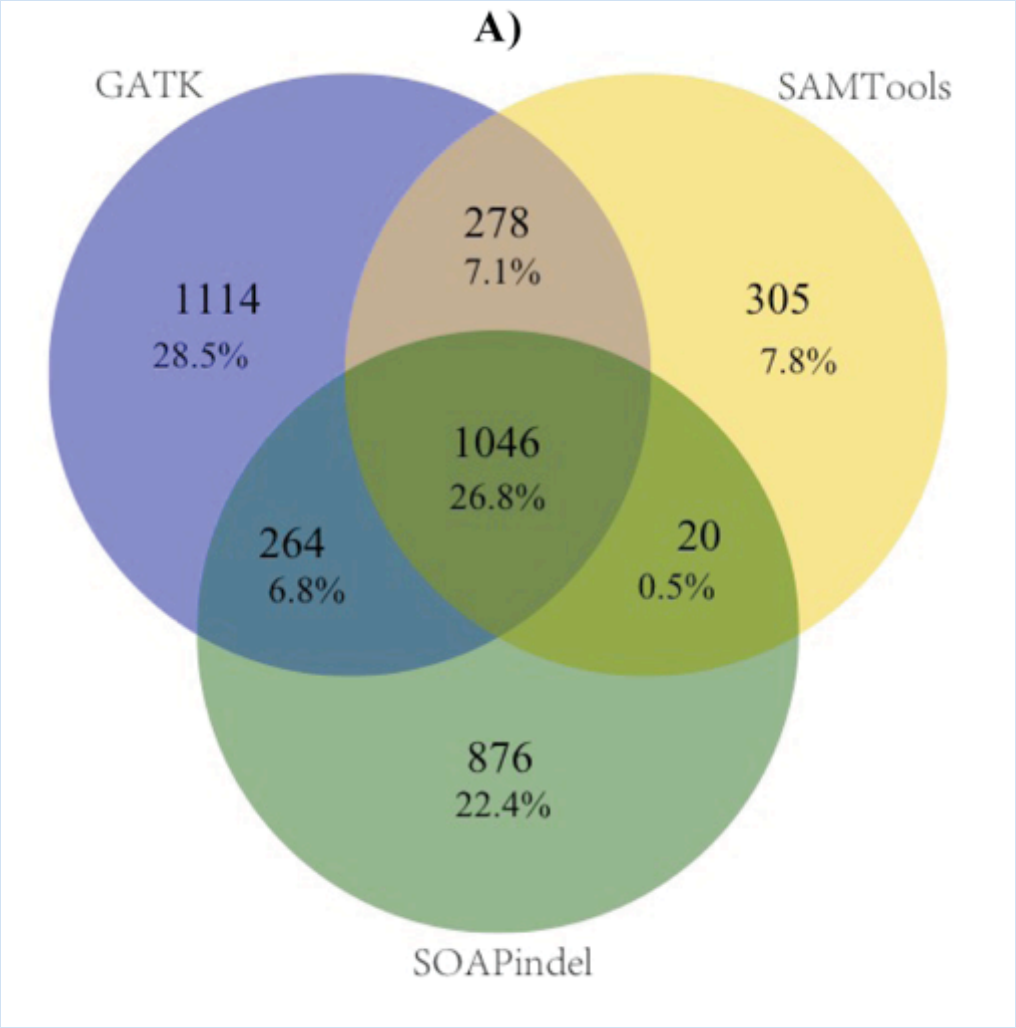




Comparing to New Versions of GATK



Validation of SNVs and Indels called by GATK, SOAP and both, with another platform

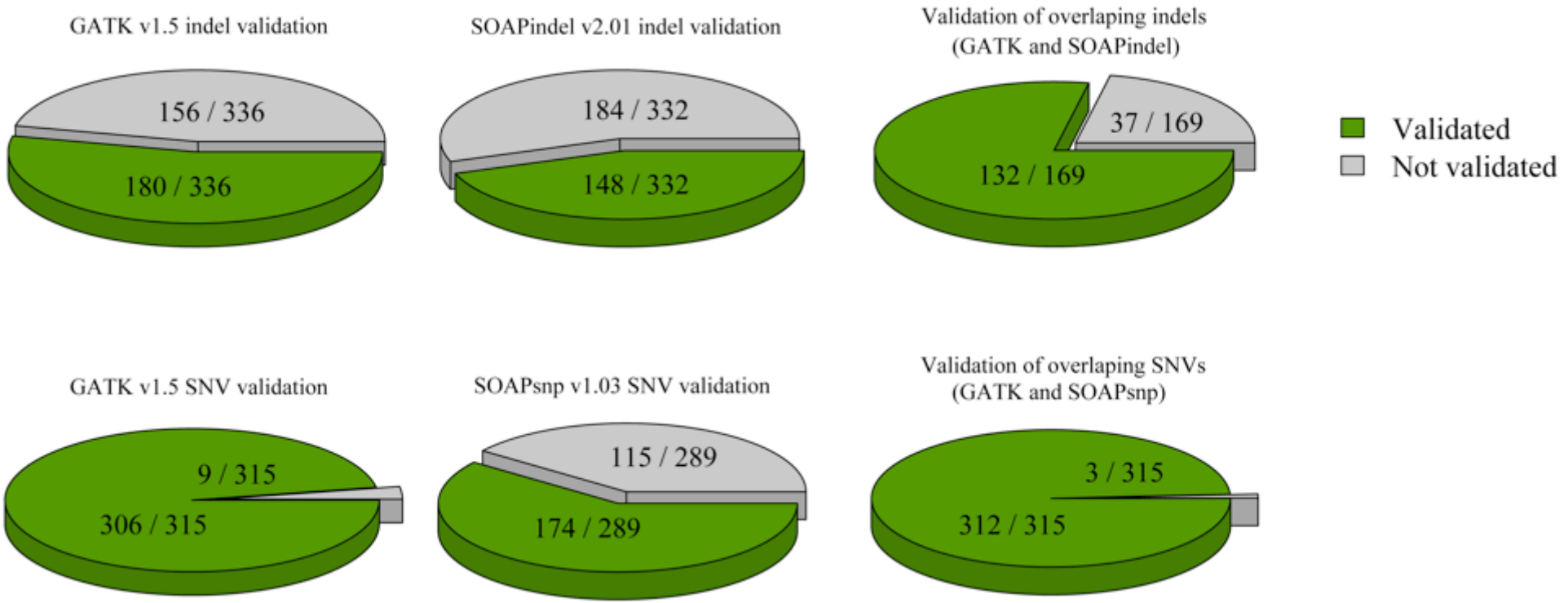


Indels

Validation with PCR amplicons and MiSeq 150 bp reads at ~5000x coverage

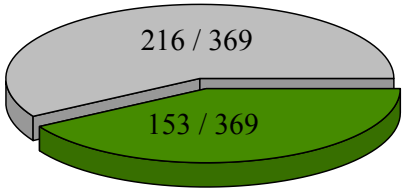
1,140 SNVs, with random sampling of 380 from the set of unique-to-GATK SNVs, 380 from the set of unique-to-SOAPsnp SNVs, and 380 from the set that were overlapping between these two pipelines.

960 indels, with random sampling of 386 from the unique-to-GATK indel set, 387 from the unique-to-SOAPindel set, and 187 from the set of indels overlapping between the two (SOAPindel and GATK).

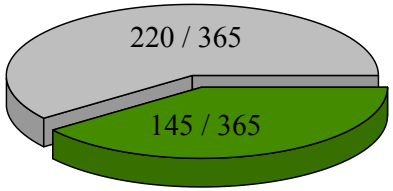


Validation of ~2000 PCR amplicons with PacBio reads from two SMRT cells (~50,000 useable reads per cell)

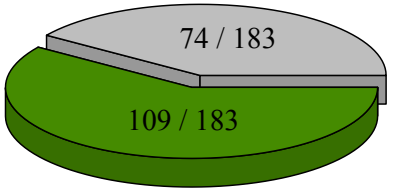
GATK v1.5 indel validation

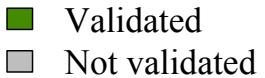


SOAPindel v2.01 indel validation

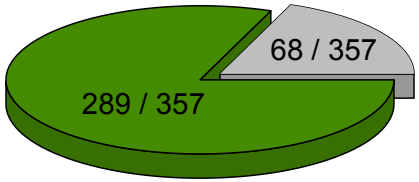


Validation of overlaping indels (GATK and SOAPindel)

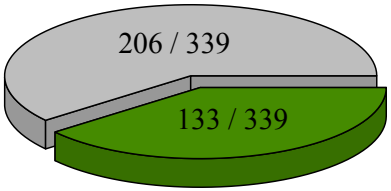




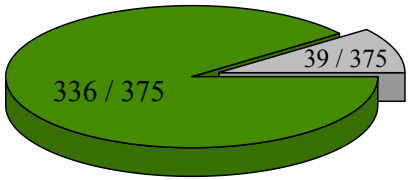
GATK v1.5 SNV validation



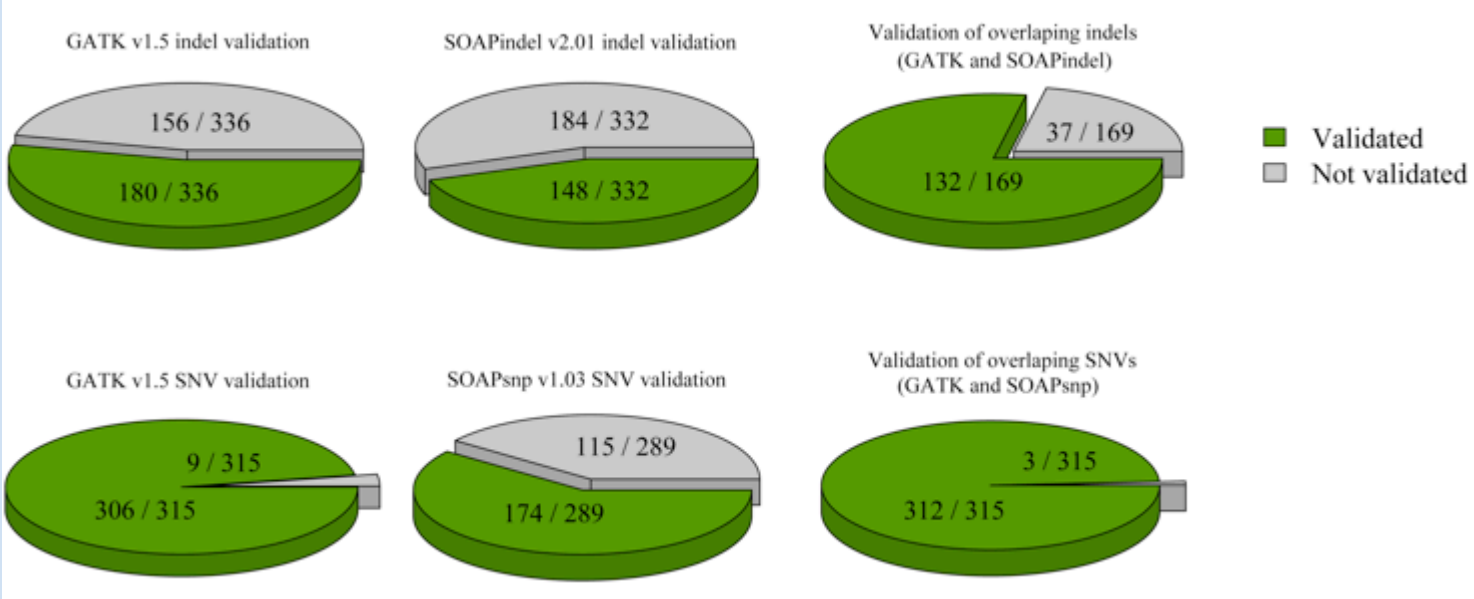
SOAPsnp v1.03 SNV validation



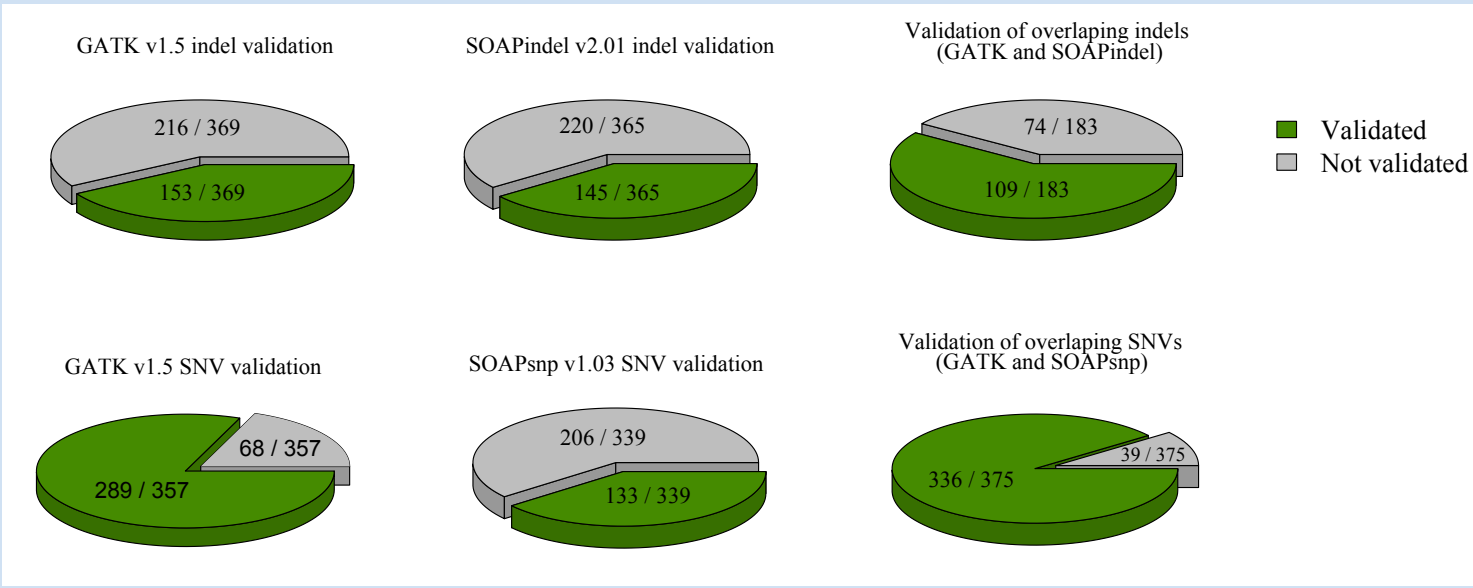
Validation of overlaping SNVs (GATK and SOAPsnp)

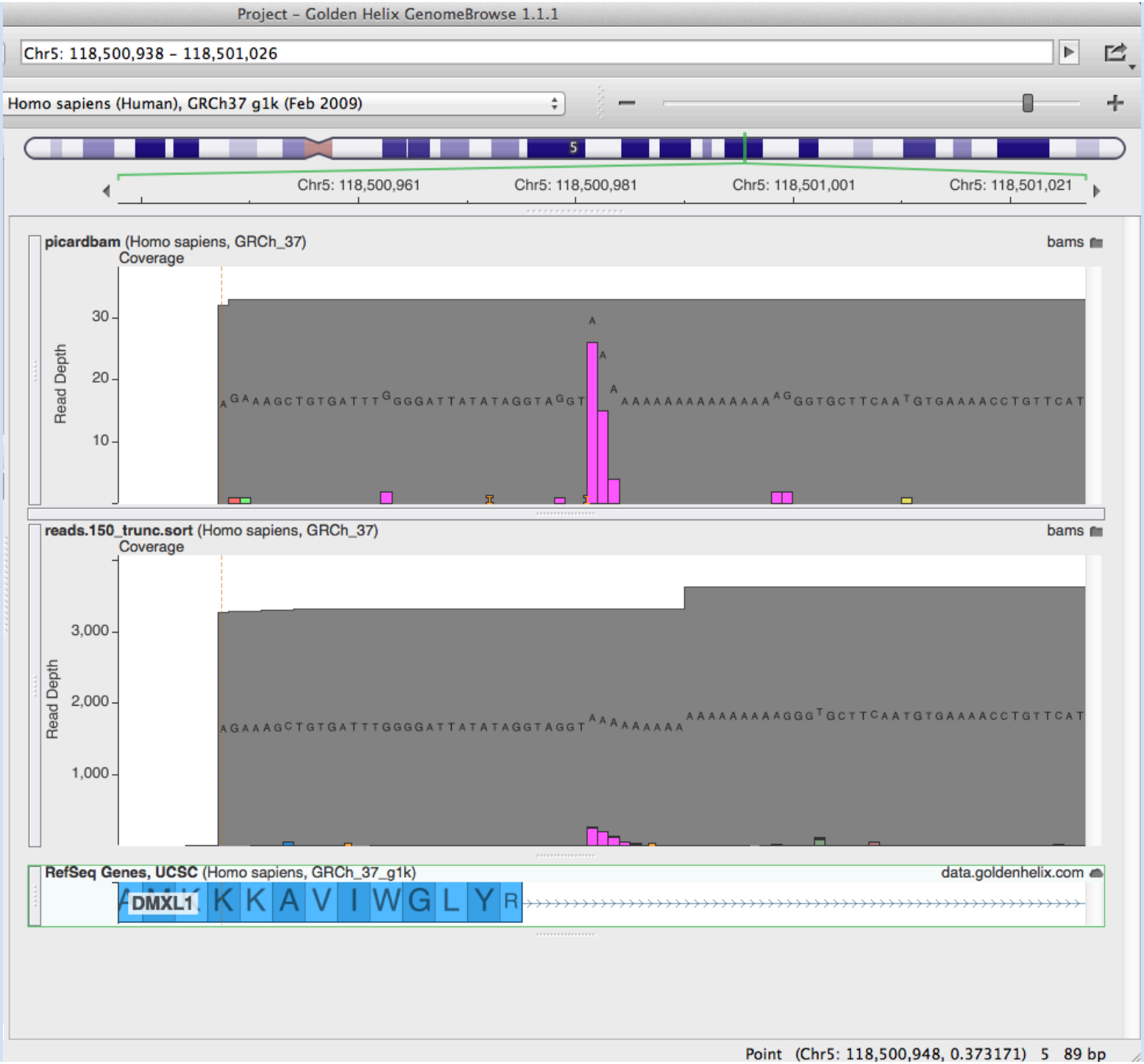


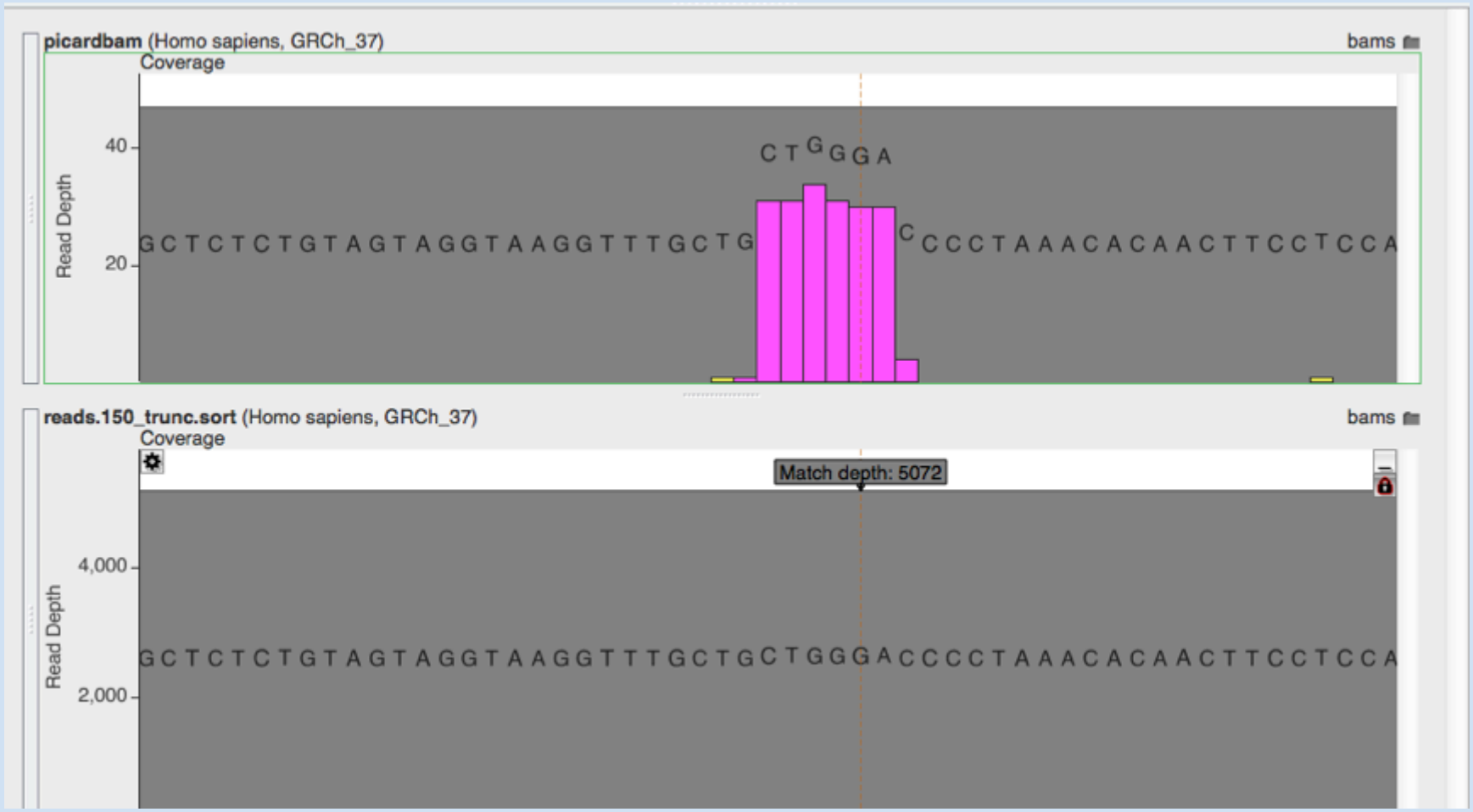
MiSeq



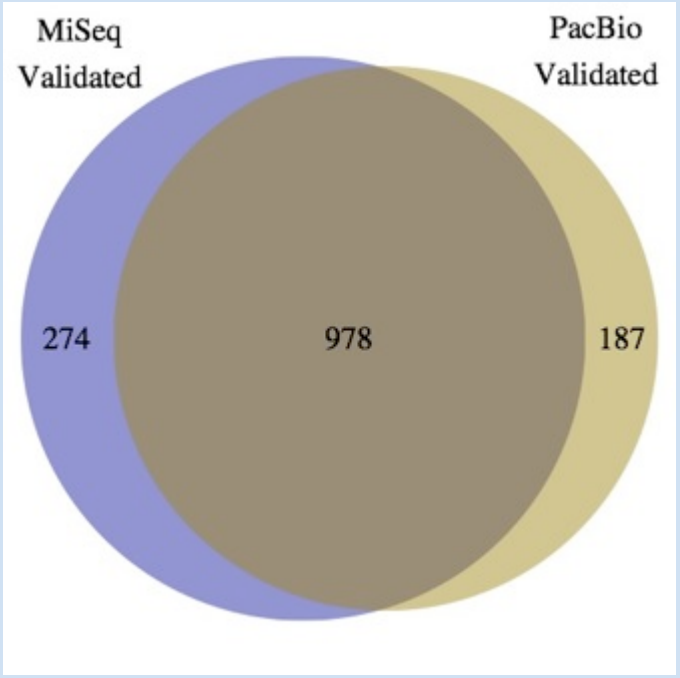
PacBio

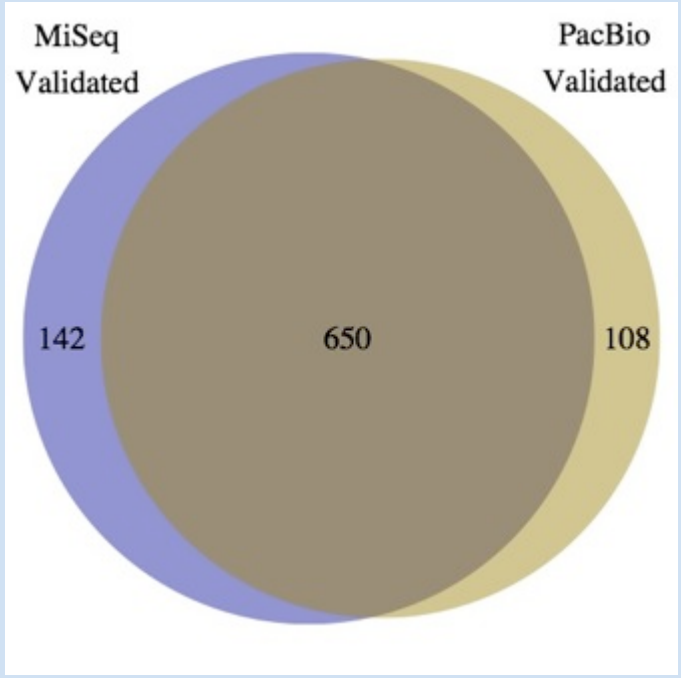



Project – Golden Helix GenomeBrowse 1.1.1
Chr5: 118,500,938 – 118,501,026
Homo sapiens (Human), GRCh37 g1k (Feb 2009)
Chr5: 118,500,961
Chr5: 118,500,981
Chr5: 118,501,001
Chr5: 118,501,021
picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
30
20
10
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
3,000
2,000
1,000
RefSeq Genes, UCSC (Homo sapiens, GRCh_37_g1k)
data.goldenhelix.com
DMXL1
K K A V I W G L Y R
Point (Chr5: 118,500,948, 0.373171) 5 89 bp

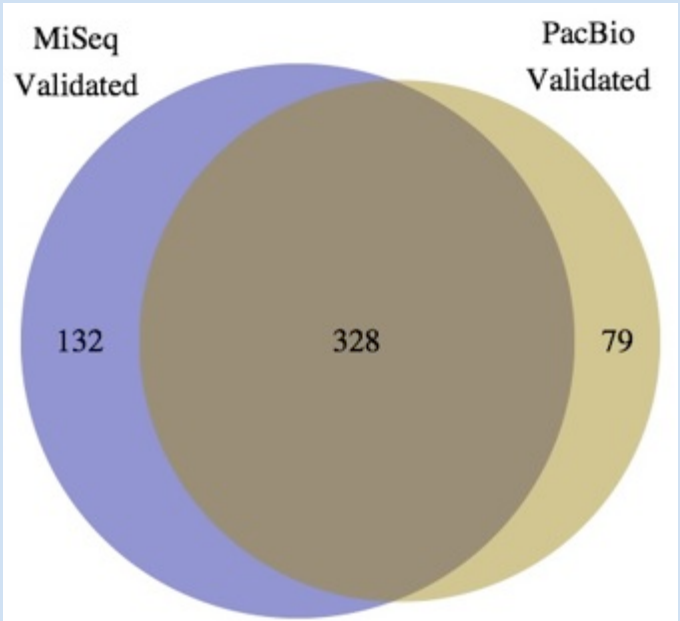

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
40
20
CTGGGA
GCTCTCTGTAGTAGGTAAGGTTTGCTG
CCCCTAAACACAACTTCCTCCA
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Match depth: 5072
Read Depth
4,000
2,000
GCTCTCTGTAGTAGGTAAGGTTTGCTGCTGGGACCCCTAAACACAACTTCCTCCA

SNVs
+ Indels

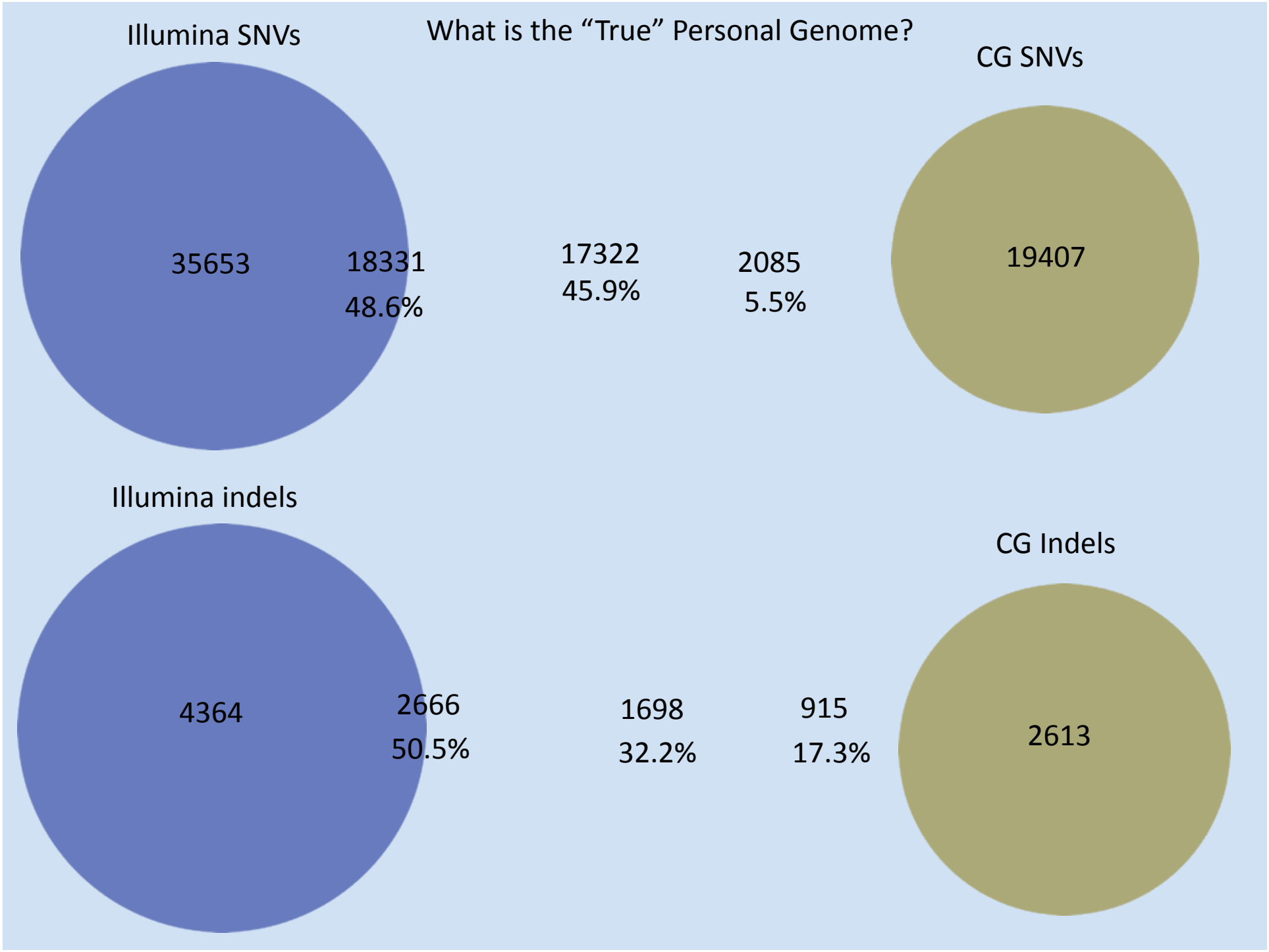


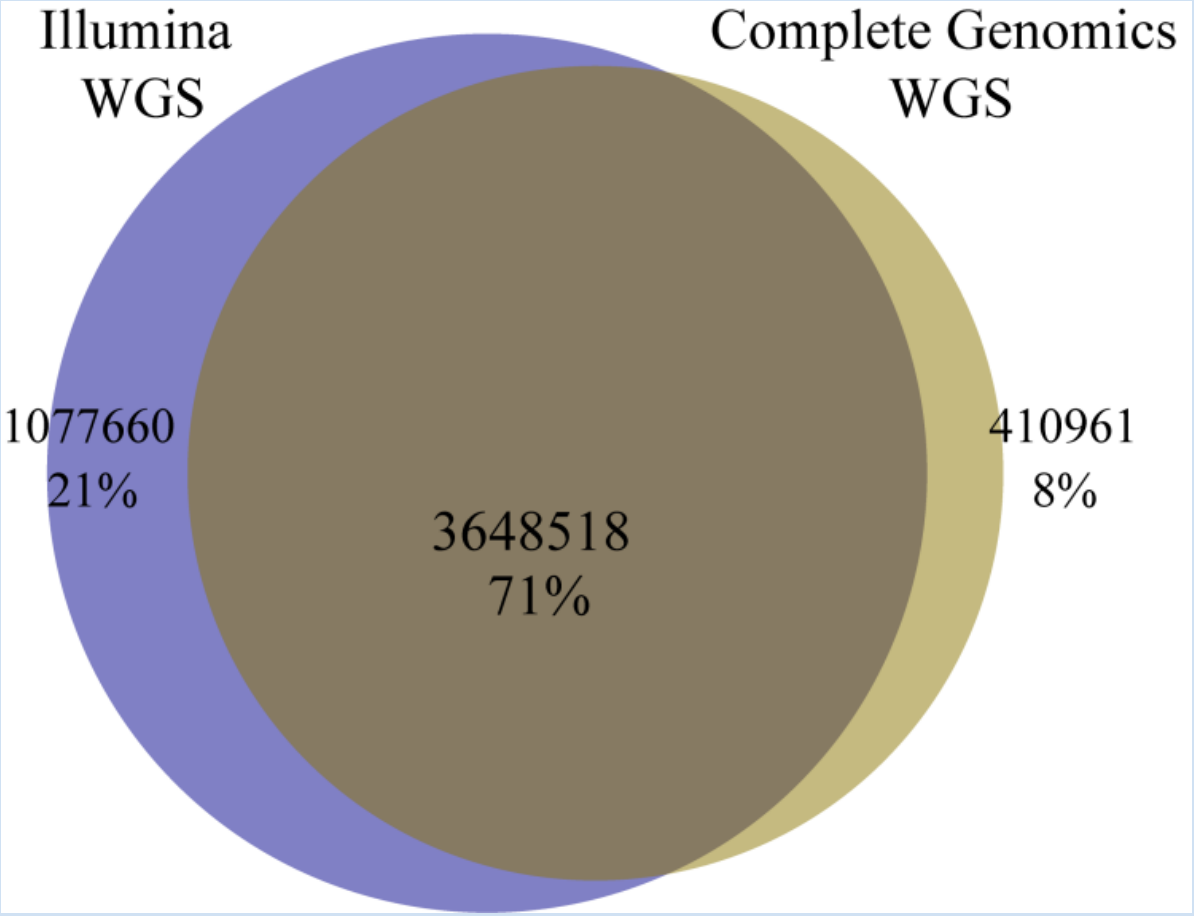


SNVs



Indels


What is the “True” Personal Genome?
Illumina SNVs
CG SNVs
35653
18331
48.6%
17322
45.9%
2085
5.5%
19407
Illumina indels
CG Indels
4364
2666
50.5%
1698
32.2%
915
17.3%
2613



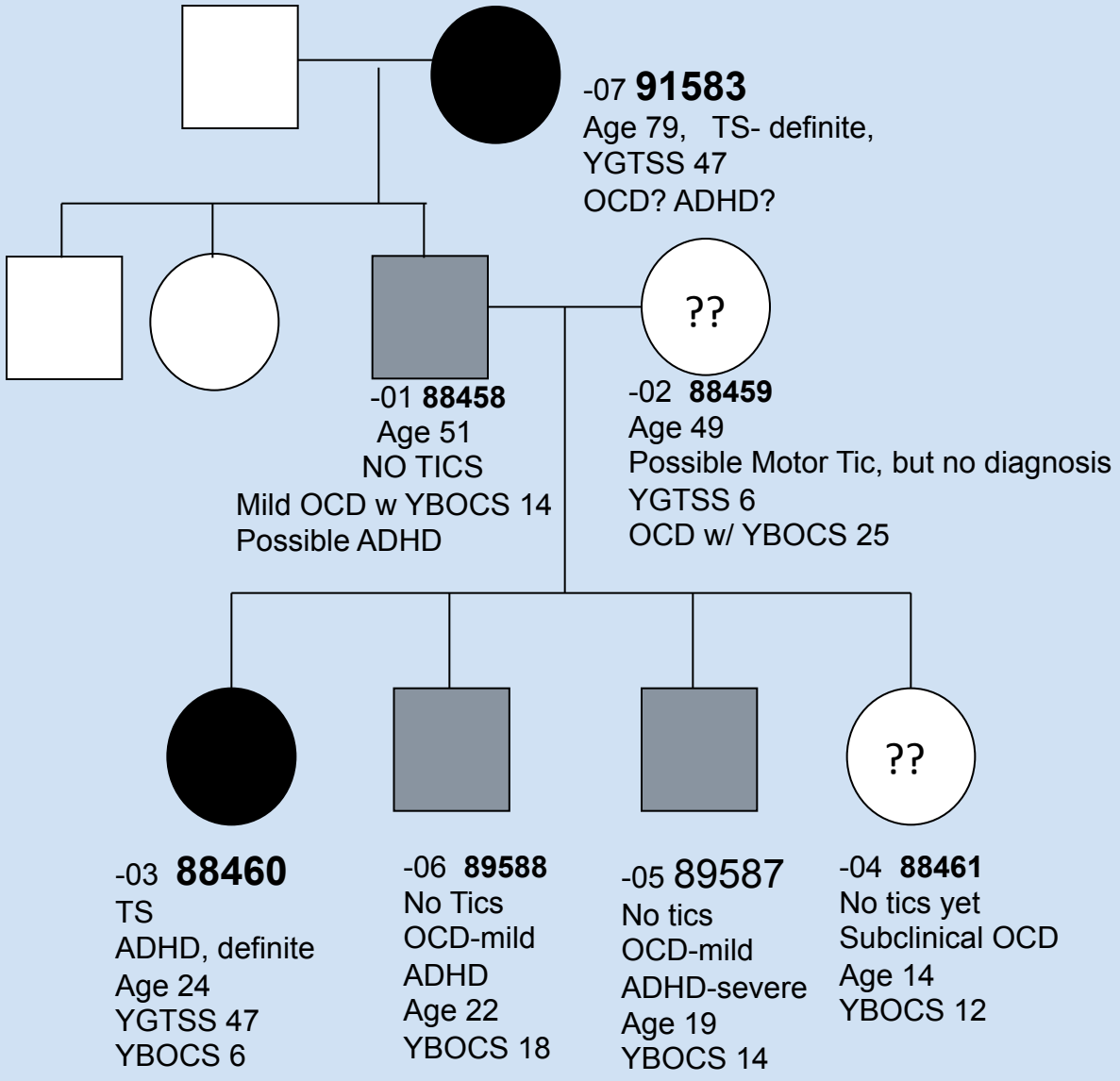
- Mean concordance across five samples between the Complete Genomics v2.0 and the Illumina HiSeq 2000 BWA/GATK whole genome sequencing and analysis pipelines is 71%.
- On average, the CG pipeline detected 410,961 variants that the Illumina BWA/GATK pipeline did not; however, the Illumina BWA/GATK pipeline detected more than double the amount of unique to pipeline variants, 1,077,660.

Optimizing the Variant Calling Pipeline Using Family Relationships

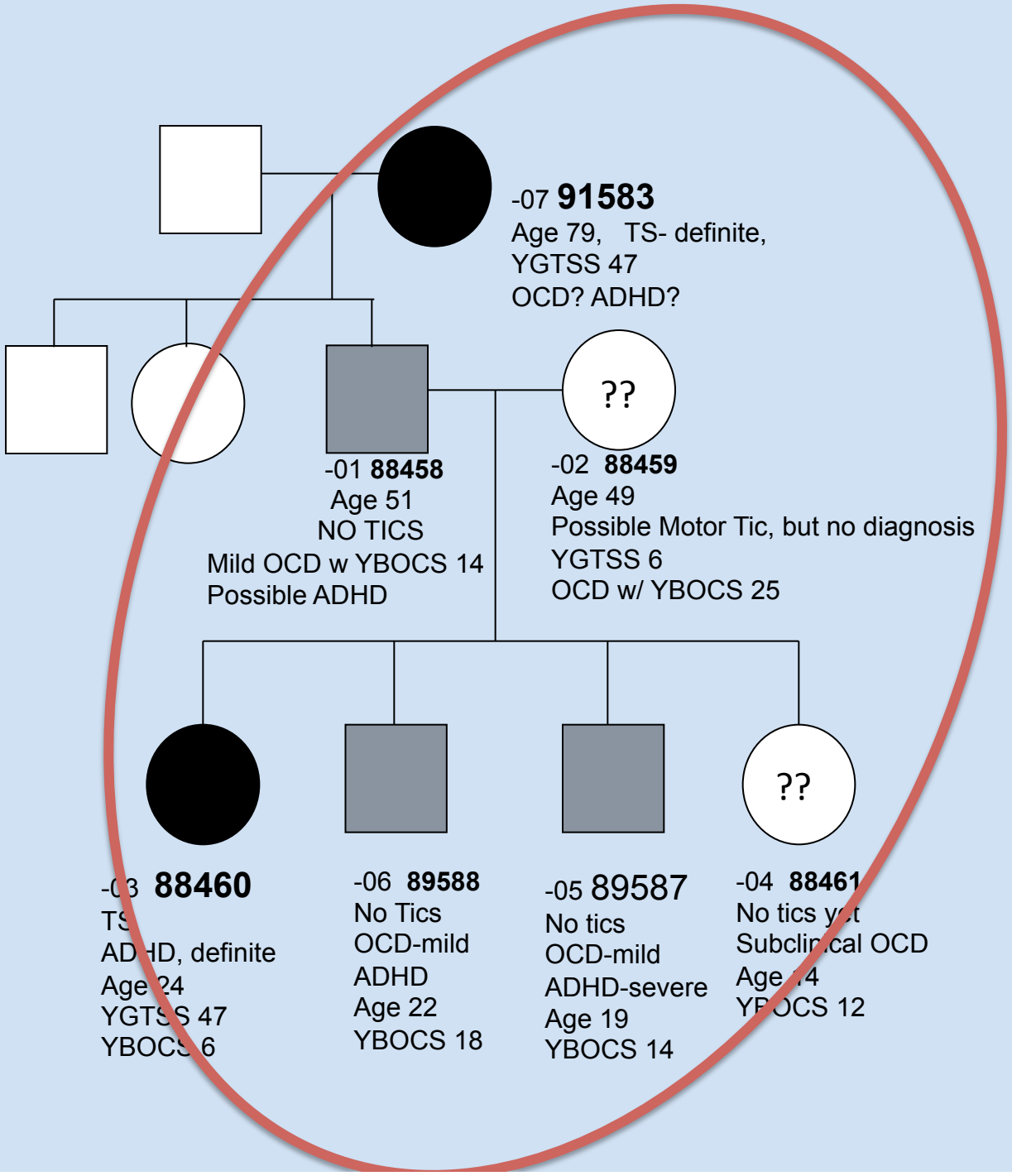
We looked for SNVs that were detected in children but not in parents using 3 different strategies:

1. We used all of the SNVs that were detected by all 5 pipelines for both parents and children
2. We used all of the detected SNVs for parents, but only the concordant SNVs between the 5 different pipelines for children.
3. We used SNVs concordant between the 5 different pipelines for children and parents.

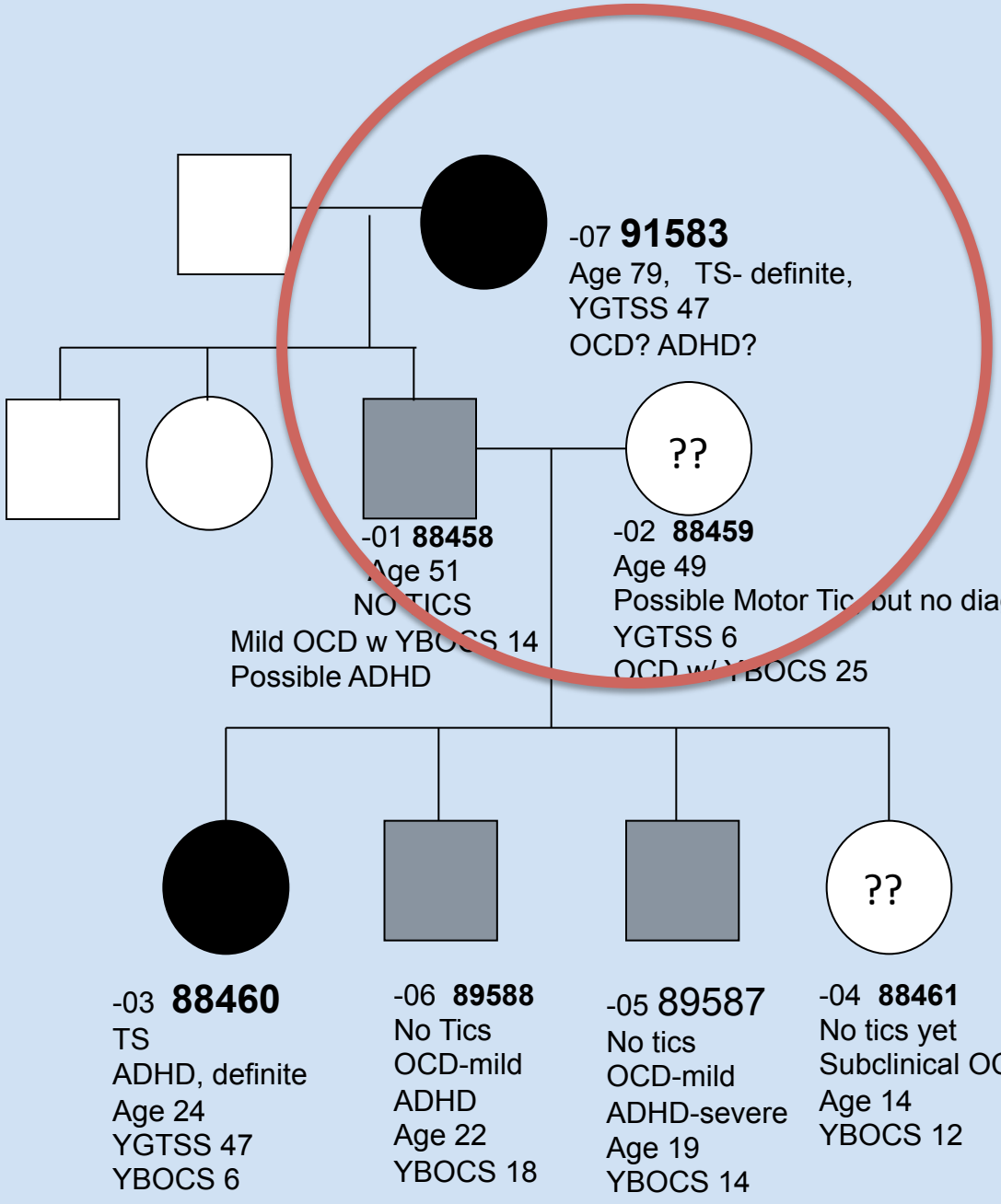
TDT- 09 -1018
K26679



TDT- 09 -1018
K26679



TDT- 09 -1018
K26679



"Parents"

Analysis based on various pipelines

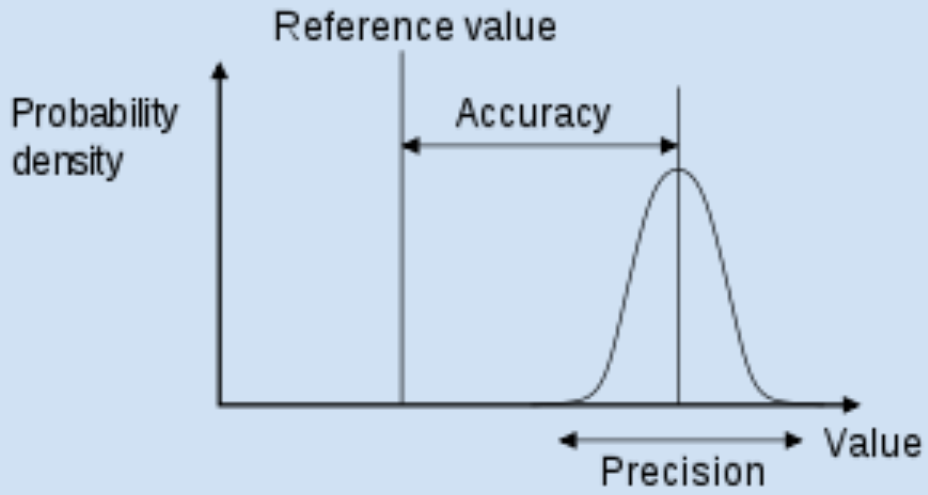
- “Parents” in this case means the mother, father AND grandmother.
- Taking the **Union** of SNVs from all 5 pipelines from “Parents”, and subtract that from the **Union** of all SNVs in each child.
- Or Subtract the **Union** of these “Parents” from the SNVs in the child **concordant** between 5 pipelines.
- Or, subtract the **concordant** variants from 5 pipelines in “Parents” from the **concordant** variants for 5 pipelines in each child.

Table 3. *De novo* single-nucleotide variants (SNVs) were detected in two families contained within the 15 study exomes.

	Number of putative *de novo* coding non-synonymous or nonsense SNVs detected	
Family 1	Without using the grandparents as a filter	Using the grandparents as a filter
Child A	241	1
Child B	211	0
Child C	102	6
Child D	242	3
Family 2		
Child A	49	NA[a]
Child B	41	NA[a]

[a]N/A, no grandparent available.

Family 1 had a grandparent available for filtering purposes, whereas family 2 did not. To minimize false positives in the pool of SNVs associated with each child, only highly concordant SNVs were used (SNVs detected by all five pipelines). To construct a comprehensive set of SNVs for each parent, and hence increase filtering accuracy, false negatives for parent SNVs were reduced by taking the union of all SNV calls from all five pipelines.



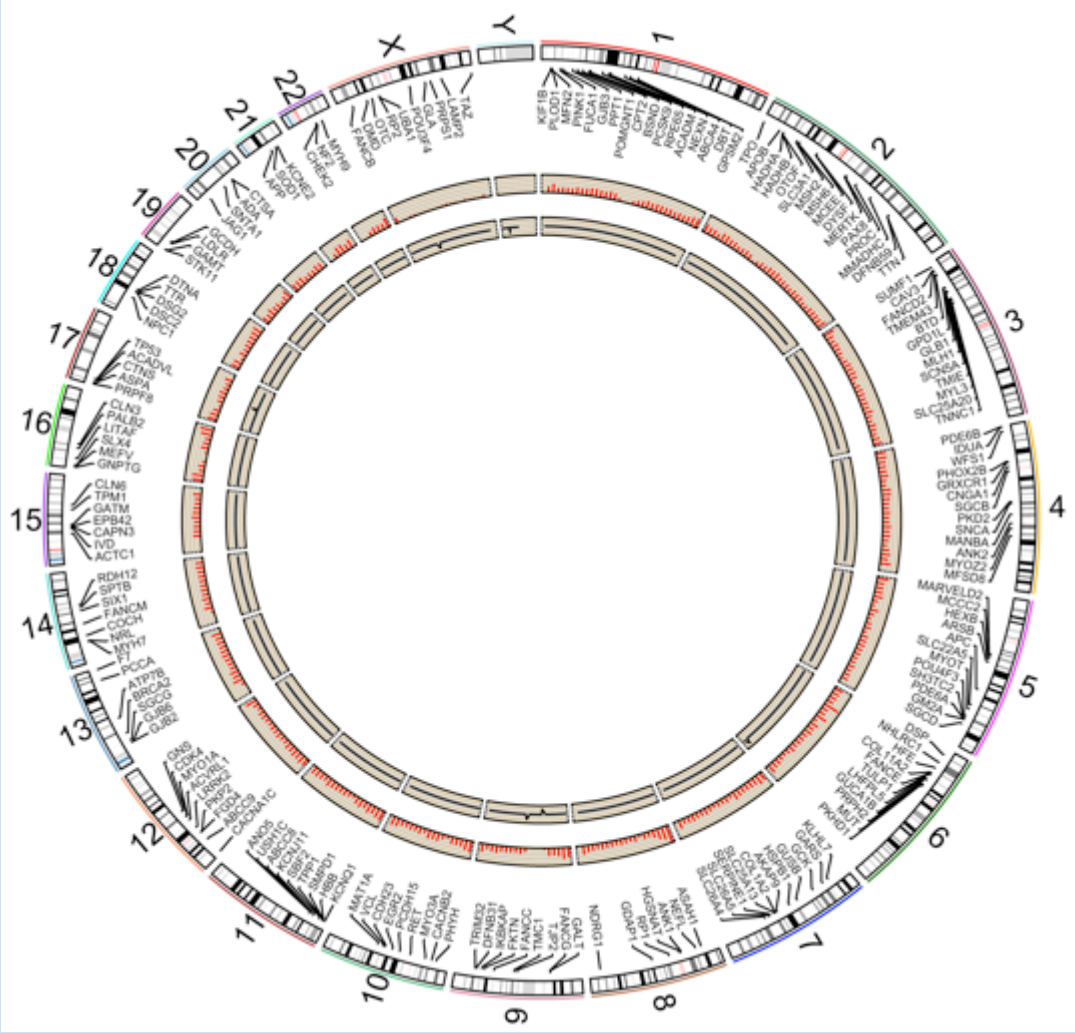
High accuracy, but low precision

High precision, but low accuracy

In the fields of science, engineering, industry, and statistics, the **accuracy** of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The **precision** of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

http://en.wikipedia.org/wiki/Accuracy_and_precision

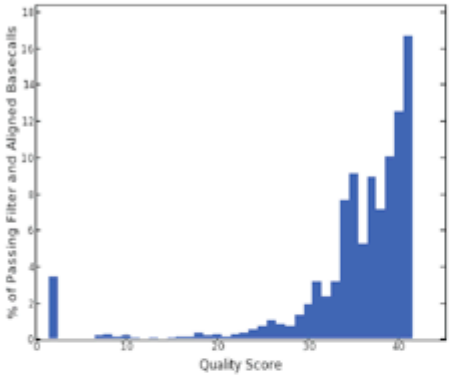
Vignette #2: One person with very severe obsessive compulsive disorder, depression and intermittent psychoses

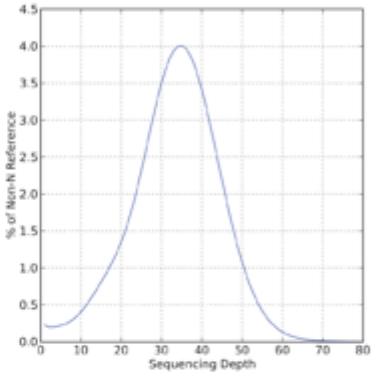


Data Volume and Quality

	Yield (Gigabases)	% Bases ≥ Q30	% Bases Aligned
Passing Filter	113.10	87.10%	87.80%

	% Callable	% ≥ 5x depth	% ≥ 10x depth	% ≥ 20x depth	Mean depth(x)
Non-N Reference	93.28%	97.57%	96.22%	88.54%	33.35



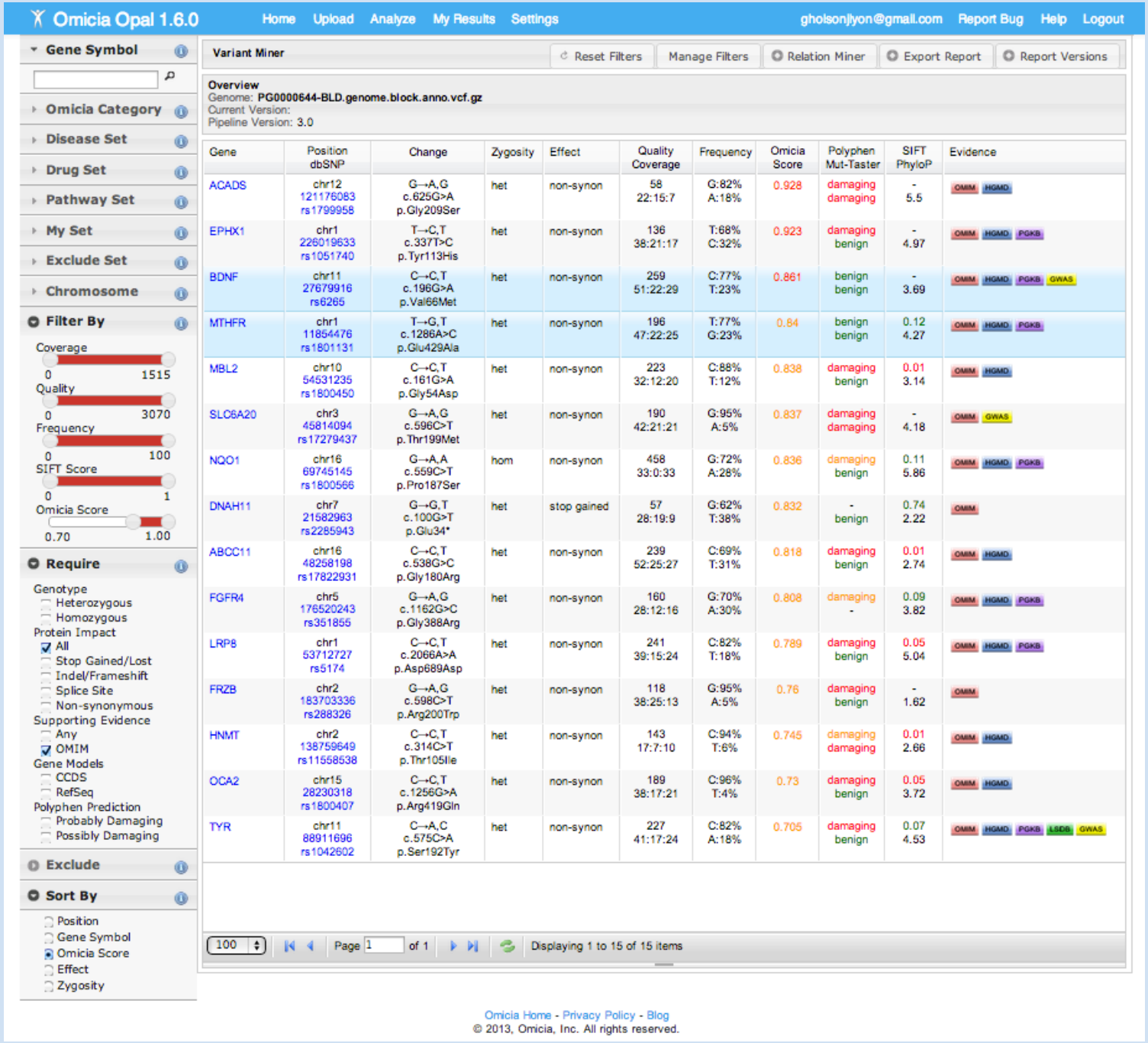


SNP Assessment

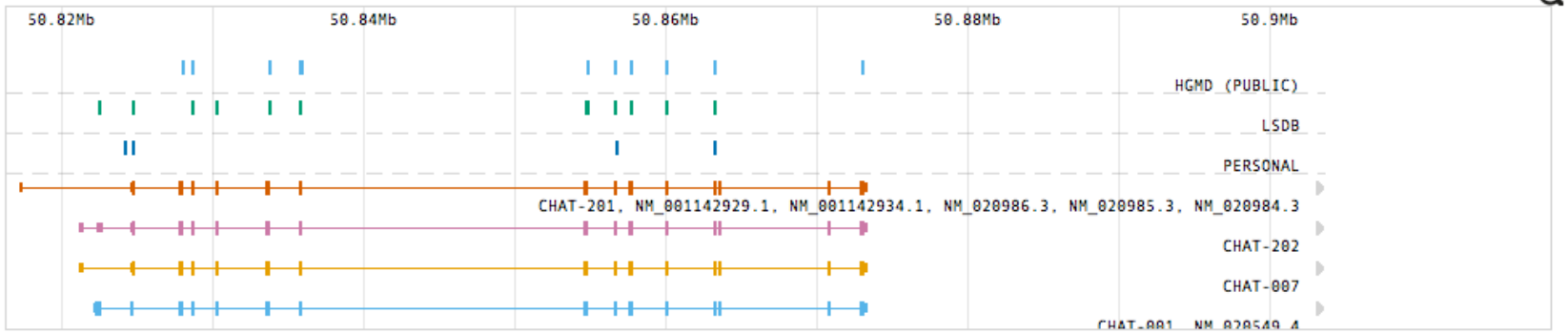
Total	Het/Hom	% in dbSNP	% in Genes	% in Coding
3,308,246	1.61	98.13%	45.47%	0.63%

Variant Statistics

	SNVs
Total Number	3,308,246
Number in Genes	1,504,121
Number in Coding Regions	20,879
Number in UTRs	24,946
Splice Site Region	2,917
Stop Gained	72
Stop Lost	16
Non-synonymous	9,884
Synonymous	10,907
Mature miRNA	36


Omicia Opal 1.6.0
Home
Upload
Analyze
My Results
Settings
gholsonjlyon@gmail.com
Report Bug
Help
Logout
Gene Symbol
Omicia Category
Disease Set
Drug Set
Pathway Set
My Set
Exclude Set
Chromosome
Filter By
Coverage
0
1515
Quality
0
3070
Frequency
0
100
SIFT Score
0
1
Omicia Score
0.70
1.00
Require
Genotype
Heterozygous
Homozygous
Protein Impact
All
Stop Gained/Lost
Indel/Frameshift
Splice Site
Non-synonymous
Supporting Evidence
Any
OMIM
Gene Models
CCDS
RefSeq
Polyphen Prediction
Probably Damaging
Possibly Damaging
Exclude
Sort By
Position
Gene Symbol
Omicia Score
Effect
Zygosity
Variant Miner
Reset Filters
Manage Filters
Relation Miner
Export Report
Report Versions
Overview
Genome: PG0000644-BLD.genome.block.anno.vcf.gz
Current Version:
Pipeline Version: 3.0
Gene
Position dbSNP
Change
Zygosity
Effect
Quality Coverage
Frequency
Omicia Score
Polyphen Mut-Taster
SIFT PhyloP
Evidence
ACADS
chr12 121176083 rs1799958
G→A,G c.625G>A p.Gly209Ser
het
non-synon
58 22:15:7
G:82% A:18%
0.928
damaging damaging
- 5.5
OMIM HGMD
EPHX1
chr1 226019633 rs1051740
T→C,T c.337T>C p.Tyr113His
het
non-synon
136 38:21:17
T:68% C:32%
0.923
damaging benign
- 4.97
OMIM HGMD PGKB
BDNF
chr11 27679916 rs6265
C→C,T c.196G>A p.Val66Met
het
non-synon
259 51:22:29
C:77% T:23%
0.861
benign benign
- 3.69
OMIM HGMD PGKB GWAS
MTHFR
chr1 11854476 rs1801131
T→G,T c.1286A>C p.Glu429Ala
het
non-synon
196 47:22:25
T:77% G:23%
0.84
benign benign
0.12 4.27
OMIM HGMD PGKB
MBL2
chr10 54531235 rs1800450
C→C,T c.161G>A p.Gly54Asp
het
non-synon
223 32:12:20
C:88% T:12%
0.838
damaging benign
0.01 3.14
OMIM HGMD
SLC6A20
chr3 45814094 rs17279437
G→A,G c.596C>T p.Thr199Met
het
non-synon
190 42:21:21
G:95% A:5%
0.837
damaging damaging
- 4.18
OMIM GWAS
NQO1
chr16 69745145 rs1800566
G→A,A c.559C>T p.Pro187Ser
hom
non-synon
458 33:0:33
G:72% A:28%
0.836
damaging benign
0.11 5.86
OMIM HGMD PGKB
DNAH11
chr7 21582963 rs2285943
G→G,T c.100G>T p.Glu34*
het
stop gained
57 28:19:9
G:62% T:38%
0.832
- benign
0.74 2.22
OMIM
ABCC11
chr16 48258198 rs17822931
C→C,T c.538G>C p.Gly180Arg
het
non-synon
239 52:25:27
C:69% T:31%
0.818
damaging benign
0.01 2.74
OMIM HGMD
FGFR4
chr5 176520243 rs351855
G→A,G c.1162G>C p.Gly388Arg
het
non-synon
160 28:12:16
G:70% A:30%
0.808
damaging -
0.09 3.82
OMIM HGMD PGKB
LRP8
chr1 53712727 rs5174
C→C,T c.2066A>A p.Asp689Asp
het
non-synon
241 39:15:24
C:82% T:18%
0.789
damaging benign
0.05 5.04
OMIM HGMD PGKB
FRZB
chr2 183703336 rs288326
G→A,G c.598C>T p.Arg200Trp
het
non-synon
118 38:25:13
G:95% A:5%
0.76
damaging benign
- 1.62
OMIM
HNMT
chr2 138759649 rs11558538
C→C,T c.314C>T p.Thr105Ile
het
non-synon
143 17:7:10
C:94% T:6%
0.745
damaging damaging
0.01 2.66
OMIM HGMD
OCA2
chr15 28230318 rs1800407
C→C,T c.1256G>A p.Arg419Gln
het
non-synon
189 38:17:21
C:96% T:4%
0.73
damaging benign
0.05 3.72
OMIM HGMD
TYR
chr11 88911696 rs1042602
C→A,C c.575C>A p.Ser192Tyr
het
non-synon
227 41:17:24
C:82% A:18%
0.705
damaging benign
0.07 4.53
OMIM HGMD PGKB LSDB GWAS
100
Page 1 of 1
Displaying 1 to 15 of 15 items
Omicia Home - Privacy Policy - Blog
© 2013, Omicia, Inc. All rights reserved.

Gene Summary for CHAT



Gene Overview	
Symbol	CHAT
Name	choline O-acetyltransferase
Location	10q11.2
Summary	This gene encodes an enzyme which catalyzes the biosynthesis of the neurotransmitter acetylcholine. This gene product is a characteristic feature of cholinergic neurons, and changes in these neurons may explain some of the symptoms of Alzheimer's disease. Polymorphisms in this gene have been associated with Alzheimer's disease and mild cognitive impairment. Mutations in this gene are associated with congenital myasthenic syndrome associated with episodic apnea. Multiple transcript variants encoding different isoforms have been found for this gene, and some of these variants have been shown to encode more than one isoform. [provided by RefSeq, May 2010]

Relevant Reference Resources	
NCBI Gene	http://www.ncbi.nlm.nih.gov/gene/1103
GeneTests	http://www.ncbi.nlm.nih.gov/sites/GeneTests/lab/gene/CHAT
Ensembl	http://www.ensembl.org/human/Gene/Summary?g=ENSG00000070748
UCSC Gene Browser	http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&db=hg19&singleSearch=knownCanonical&position=CHAT
Genetics Home Reference	http://ghr.nlm.nih.gov/gene/CHAT

Associated Disease Categories

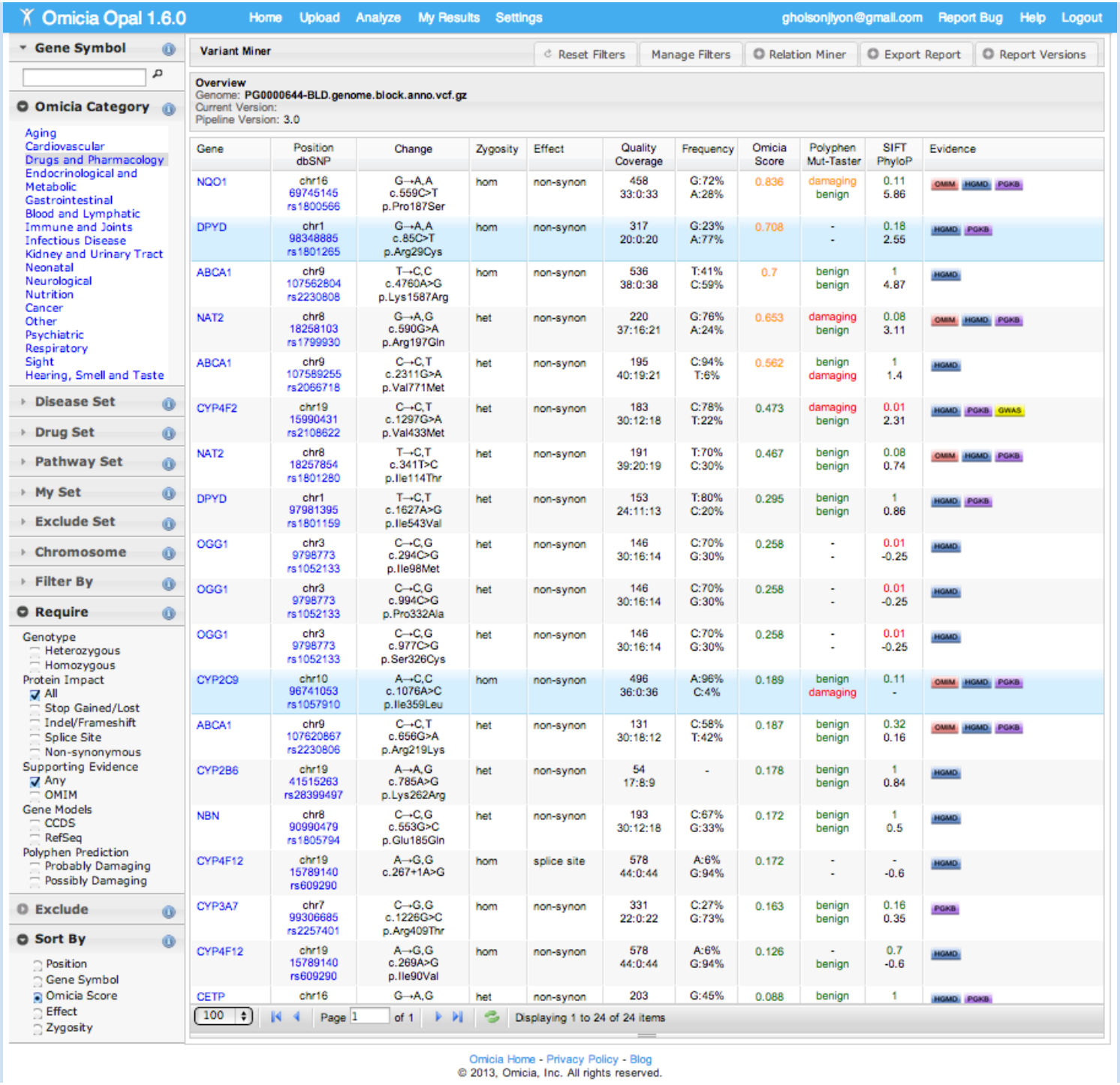
Category	Disease	Citation
DRUGS, CLINICAL PHARMACOLOGY AND ENVIRONMENT	Drug toxicity	Roden et al., 2002

Associated Knowledge Sets

Name	Type	Description
ODG - Alzheimers	disease	Omicia Disease Genes (ODG) Top 10 Neurological - Alzheimers
TruSight Exome	disease	Illumina's targeted rare genetic conditions exome test containing 2,761 genes covered in the HGMD database.
MitoGO	myset	
Longo - Phenomizer Fatty Acid Big	myset	A list of genes from phenomizer build from Patient Features HP:0004359. Long List ~3000 genes

Personal Variants in this Gene

Position	Transcript	Transcript HGVS	Protein	Protein HGVS	Zyg	Effect
50824117	NM_001142933.1	c.19G>A	NP_001136405	p.Asp7Asn	het	non-synon
50824619	NM_001142933.1	c.112G>A	NP_001136405	p.Ala38Thr	het	non-synon
50856652	NM_020549	c.1382G>A	NP_065574	p.Val461Met	hom	non-synon
50863147	NM_020549	c.1642T>C	NP_065574	p.His548His	hom	synonymous


Omicia Opal 1.6.0
Home
Upload
Analyze
My Results
Settings
gholsonjlyon@gmail.com
Report Bug
Help
Logout
Gene Symbol
Omicia Category
Aging
Cardiovascular
Drugs and Pharmacology
Endocrinological and Metabolic
Gastrointestinal
Blood and Lymphatic
Immune and Joints
Infectious Disease
Kidney and Urinary Tract
Neonatal
Neurological
Nutrition
Cancer
Other
Psychiatric
Respiratory
Sight
Hearing, Smell and Taste
Disease Set
Drug Set
Pathway Set
My Set
Exclude Set
Chromosome
Filter By
Require
Genotype
Heterozygous
Homozygous
Protein Impact
All
Stop Gained/Lost
Indel/Frameshift
Splice Site
Non-synonymous
Supporting Evidence
Any
OMIM
Gene Models
CCDS
RefSeq
Polyphen Prediction
Probably Damaging
Possibly Damaging
Exclude
Sort By
Position
Gene Symbol
Omicia Score
Effect
Zygosity
Variant Miner
Reset Filters
Manage Filters
Relation Miner
Export Report
Report Versions
Overview
Genome: PG0000644-BLD.genome.block.anno.vcf.gz
Current Version:
Pipeline Version: 3.0
Gene
Position dbSNP
Change
Zygosity
Effect
Quality Coverage
Frequency
Omicia Score
Polyphen Mut-Taster
SIFT PhyloP
Evidence
NQO1 chr16 69745145 rs1800566 G→A,A c.559C>T p.Pro187Ser hom non-synon 458 33:0:33 G:72% A:28% 0.836 damaging benign 0.11 5.86 OMIM HGMD PGKB
DPYD chr1 98348885 rs1801265 G→A,A c.85C>T p.Arg29Cys hom non-synon 317 20:0:20 G:23% A:77% 0.708 - - 0.18 2.55 HGMD PGKB
ABCA1 chr9 107562804 rs2230808 T→C,C c.4760A>G p.Lys1587Arg hom non-synon 536 38:0:38 T:41% C:59% 0.7 benign benign 1 4.87 HGMD
NAT2 chr8 18258103 rs1799930 G→A,G c.590G>A p.Arg197Gln het non-synon 220 37:16:21 G:76% A:24% 0.653 damaging benign 0.08 3.11 OMIM HGMD PGKB
ABCA1 chr9 107589255 rs2066718 C→C,T c.2311G>A p.Val771Met het non-synon 195 40:19:21 C:94% T:6% 0.562 benign damaging 1 1.4 HGMD
CYP4F2 chr19 15990431 rs2108622 C→C,T c.1297G>A p.Val433Met het non-synon 183 30:12:18 C:78% T:22% 0.473 damaging benign 0.01 2.31 HGMD PGKB GWAS
NAT2 chr8 18257854 rs1801280 T→C,T c.341T>C p.Ile114Thr het non-synon 191 39:20:19 T:70% C:30% 0.467 benign benign 0.08 0.74 OMIM HGMD PGKB
DPYD chr1 97981395 rs1801159 T→C,T c.1627A>G p.Ile543Val het non-synon 153 24:11:13 T:80% C:20% 0.295 benign benign 1 0.86 HGMD PGKB
OGG1 chr3 9798773 rs1052133 C→C,G c.294C>G p.Ile98Met het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
OGG1 chr3 9798773 rs1052133 C→C,G c.994C>G p.Pro332Ala het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
OGG1 chr3 9798773 rs1052133 C→C,G c.977C>G p.Ser326Cys het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
CYP2C9 chr10 96741053 rs1057910 A→C,C c.1076A>C p.Ile359Leu hom non-synon 496 36:0:36 A:96% C:4% 0.189 benign damaging 0.11 - OMIM HGMD PGKB
ABCA1 chr9 107620867 rs2230806 C→C,T c.656G>A p.Arg219Lys het non-synon 131 30:18:12 C:58% T:42% 0.187 benign benign 0.32 0.16 OMIM HGMD PGKB
CYP2B6 chr19 41515263 rs28399497 A→A,G c.785A>G p.Lys262Arg het non-synon 54 17:8:9 - 0.178 benign benign 1 0.84 HGMD
NBN chr8 90990479 rs1805794 C→C,G c.553G>C p.Glu185Gln het non-synon 193 30:12:18 C:67% G:33% 0.172 benign benign 1 0.5 HGMD
CYP4F12 chr19 15789140 rs609290 A→G,G c.267+1A>G hom splice site 578 44:0:44 A:6% G:94% 0.172 - - - -0.6 HGMD
CYP3A7 chr7 99306685 rs2257401 C→G,G c.1226G>C p.Arg409Thr hom non-synon 331 22:0:22 C:27% G:73% 0.163 benign benign 0.16 0.35 PGKB
CYP4F12 chr19 15789140 rs609290 A→G,G c.269A>G p.Ile90Val hom non-synon 578 44:0:44 A:6% G:94% 0.126 - benign 0.7 -0.6 HGMD
CETP chr16 G→A,G het non-synon 203 G:45% 0.088 benign 1 HGMD PGKB
100
Page 1 of 1
Displaying 1 to 24 of 24 items
Omicia Home - Privacy Policy - Blog
© 2013, Omicia, Inc. All rights reserved.

One person with very severe obsessive compulsive disordere, depression and intermittent psychoses

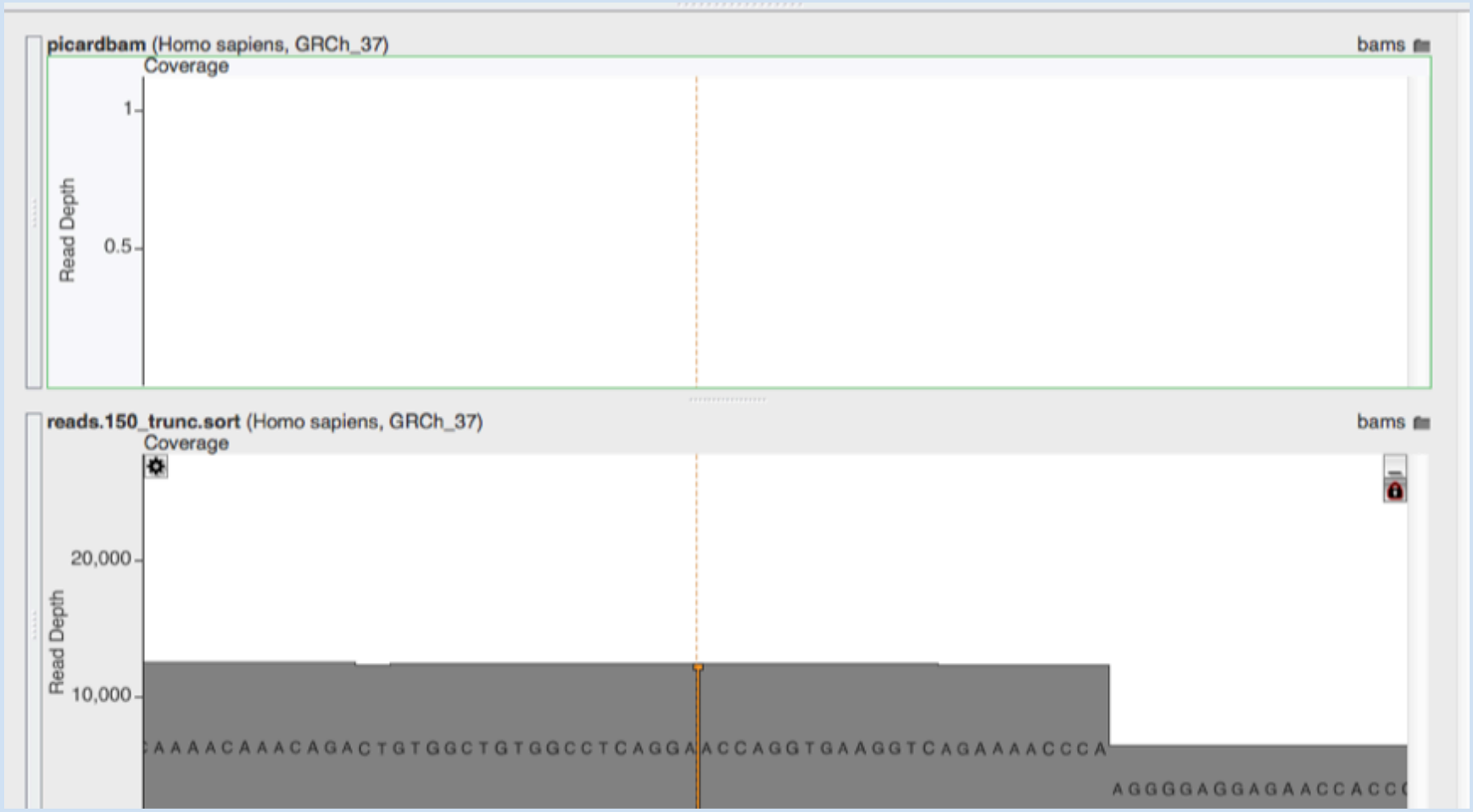
Gene name	Genomic coordinates	Amino acid change	Zygosity	Mutation type	Population Frequency	Clinical significance
MTHFR	chr1: 11854476	Glu>Ala	heterozygous	non-synon	T:77% G:23%	Susceptibility to psychoses, schizophrenia, occlusive vascular disease, neural tube defects, colon cancer, acute leukemia, and methylenetetra-hydrofolate reductase deficiency
BDNF	chr11: 27679916	Val>Met	heterozygous	non-synon	C:77% T:23%	Susceptibility to OCD, psychosis, and diminished response to exposure therapy
CHAT	chr10: 50824117	Asp>Asn	heterozygous	non-synon	G:85% A:15%	Susceptibility to schizophrenia and other psychopathological disorders.

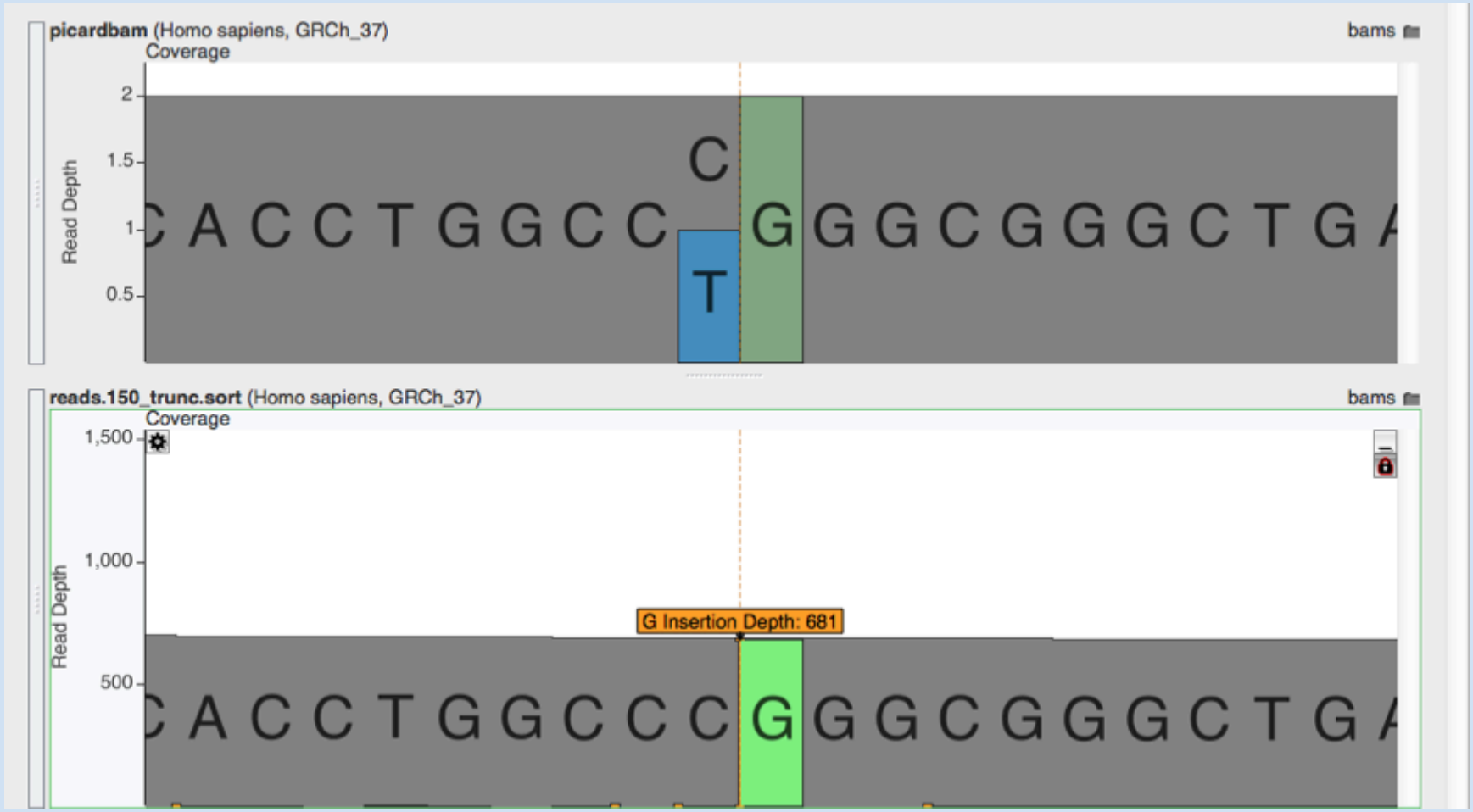
Conclusions

- Sequencing a grandparent seems to help eliminate errors derived from the current depth of sequencing coverage in the mother and father.
- For now, we advocate using more than one pipeline on one set of sequencing data, but we expect the field to move toward >2 sequencing platforms per sample.
- Still need substantial work on indel-calling and validation.

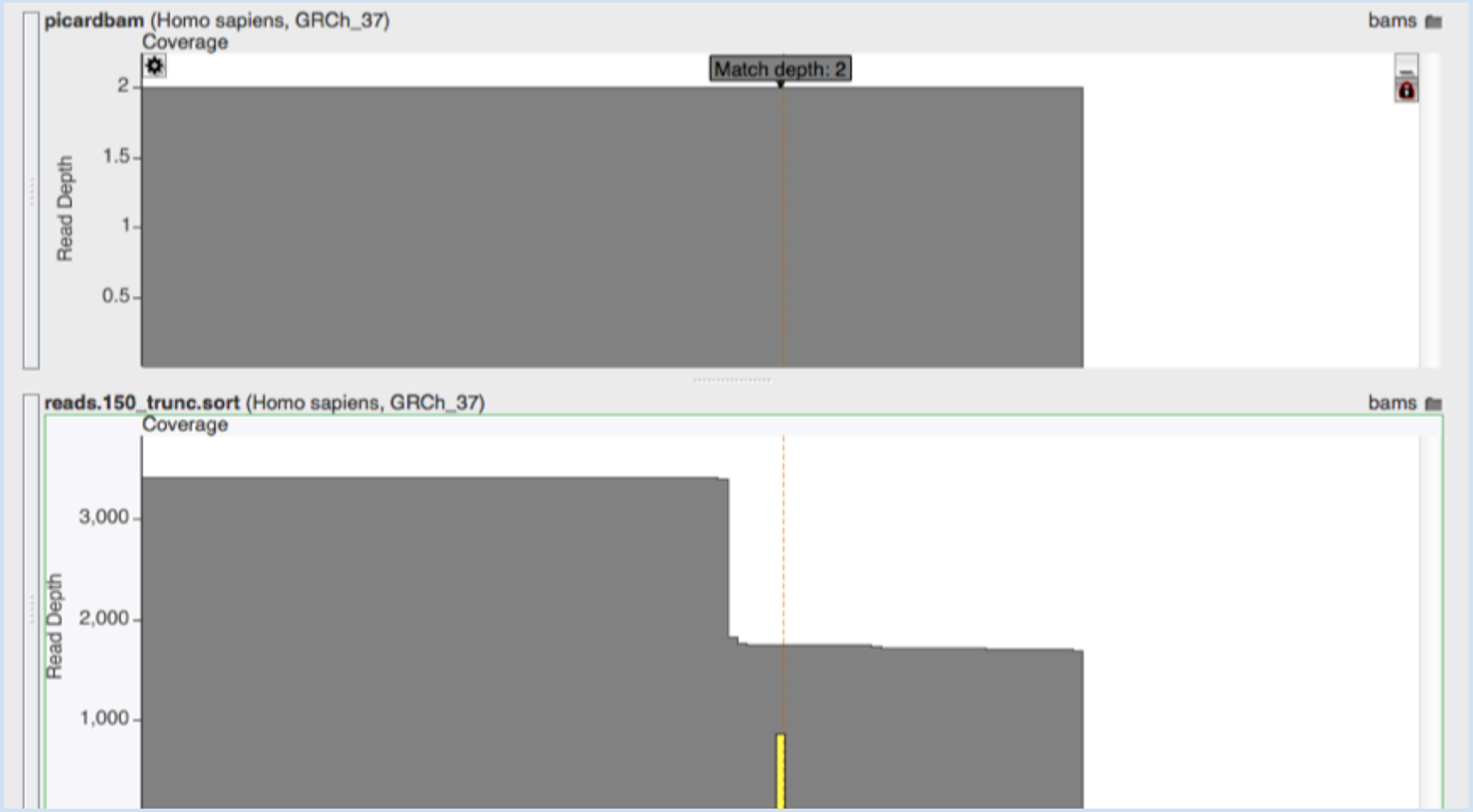
The End

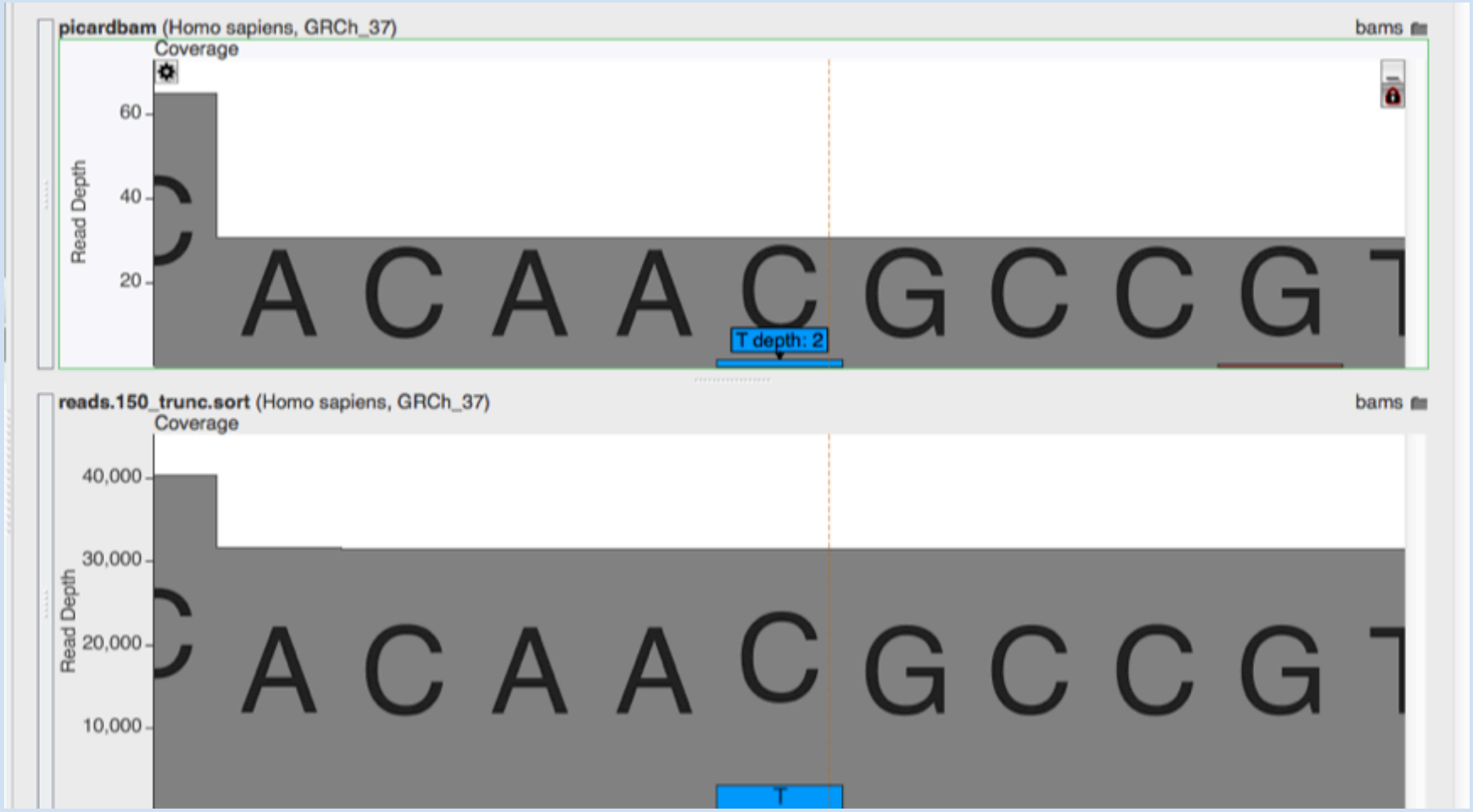
EXTRA SLIDES – Not Shown


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
1
0.5
Read Depth
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
20,000
10,000
Read Depth
CAAAACAAACAGACTGTGGCTGTGGCCTCAGGAACCAGGTGAAGGTCAGAAAACCCA
AGGGGAGGAGAACCACCC

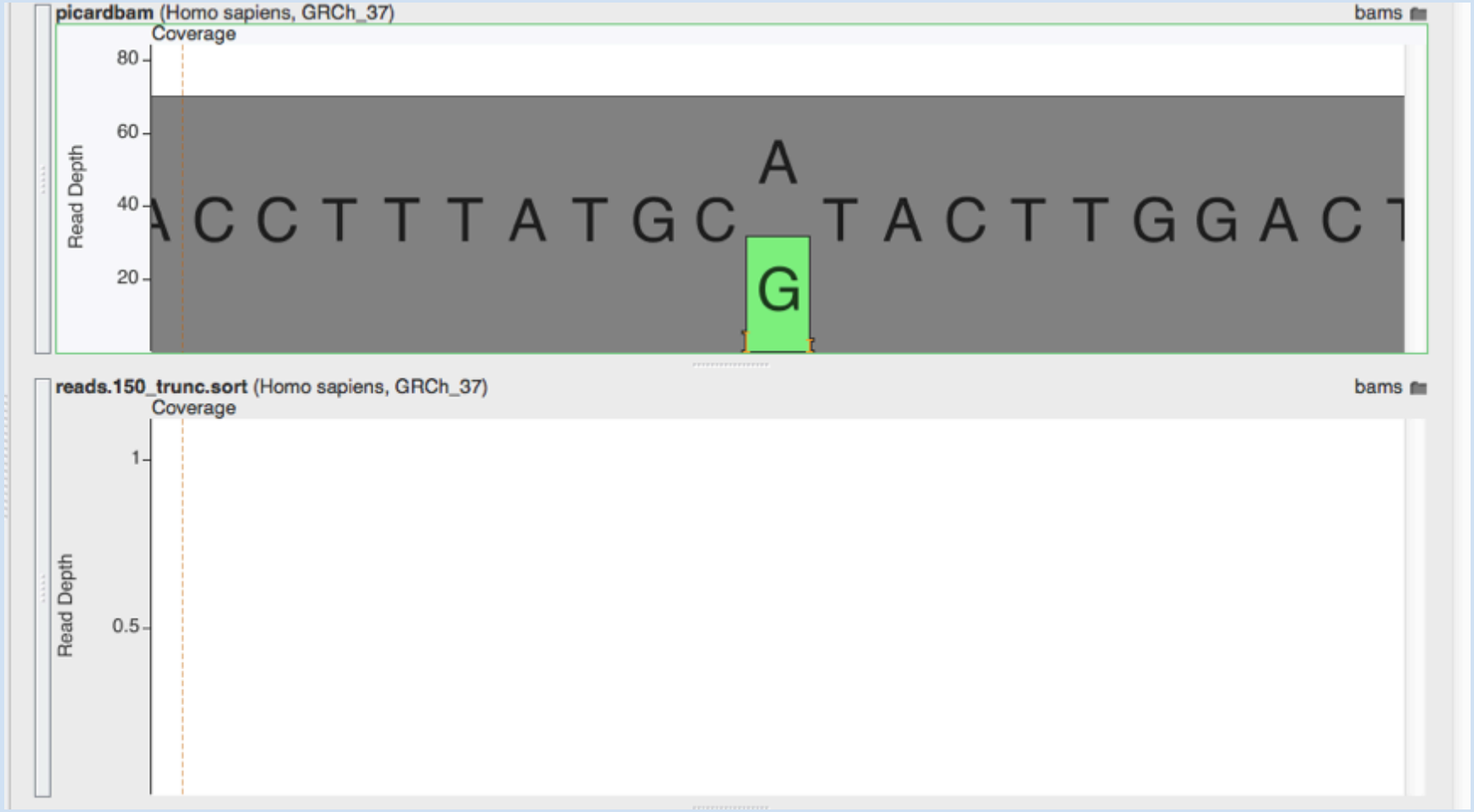

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
2
1.5
1
0.5
C
T
C A C C T G G C C G G G C G G G C T G A
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1,500
1,000
500
G Insertion Depth: 681
C A C C T G G C C C G G G C G G G C T G A


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
4
3
2
1
A
T
AATGGCTACATAGCGATCCAATGCCATTGAAGCAAGGAGTAAAACTCAGTGAT
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
6,000
4,000
2,000
AAA
CATTGAAG
TGGCTACATAGCGATCCAATGC
CAAGGAAGTAAAACTCAGTGAT


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Match depth: 2
Read Depth
2
1.5
1
0.5
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
3,000
2,000
1,000

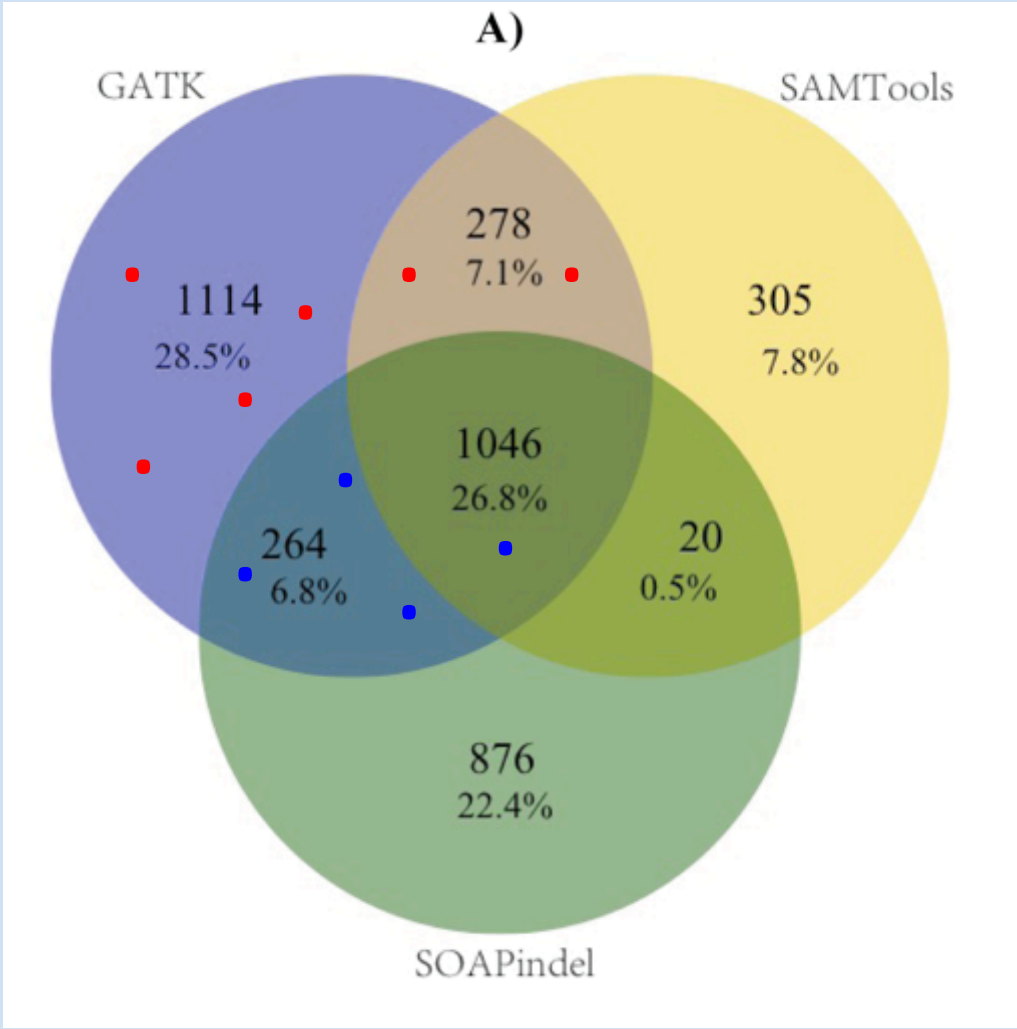

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
60
40
20
A C A A C G C C G T
T depth: 2
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
40,000
30,000
20,000
10,000
A C A A C G C C G T
T


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
5
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1
0.5

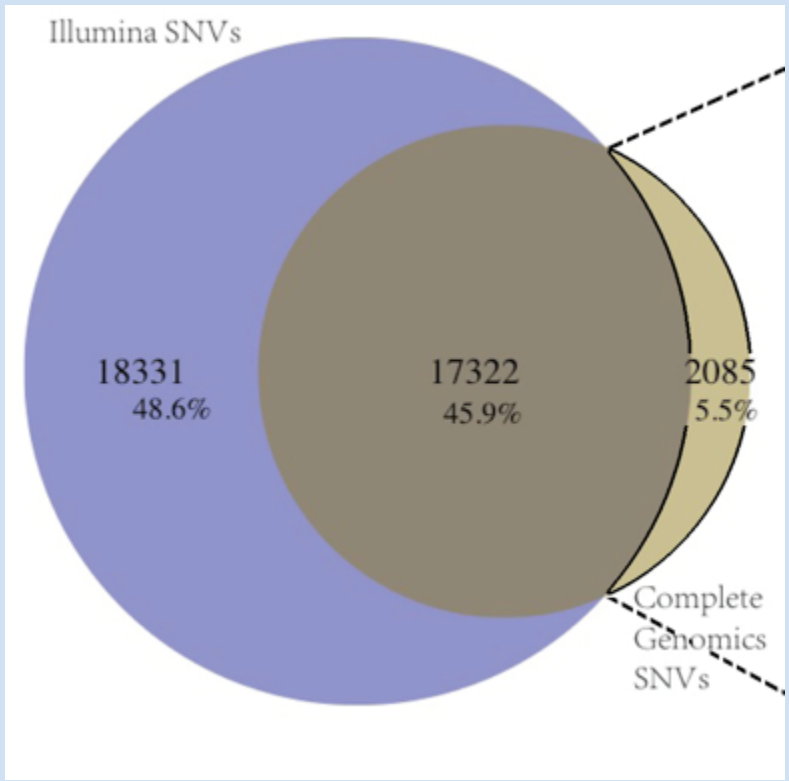

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
80
60
40
20
A
C C T T T A T G C
T A C T T G G A C
G
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1
0.5

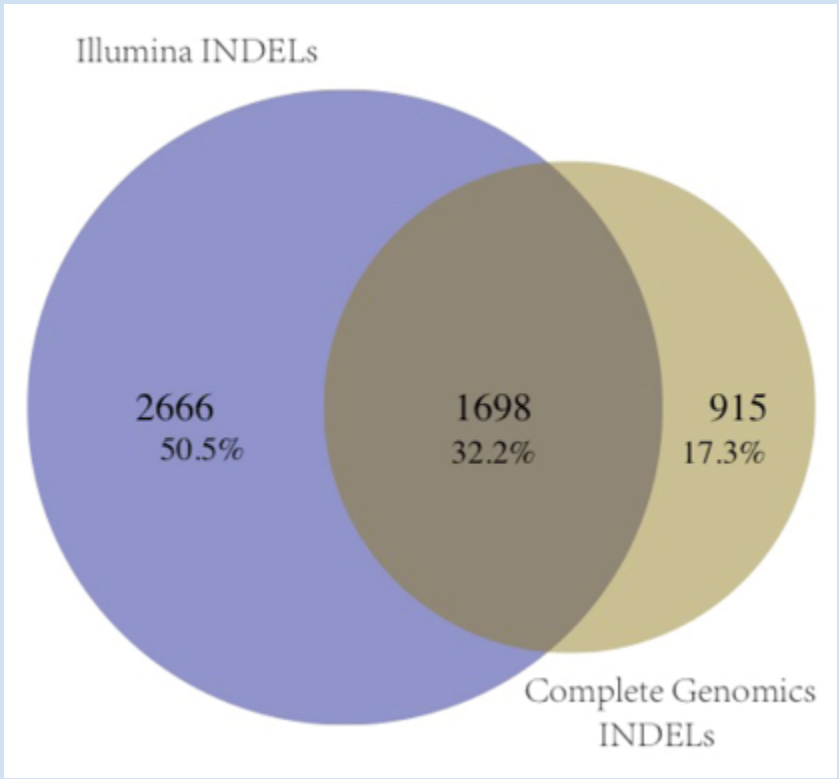

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
15
10
5
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
600
400
200

Indels called by GATK, SOAP and SAMtools

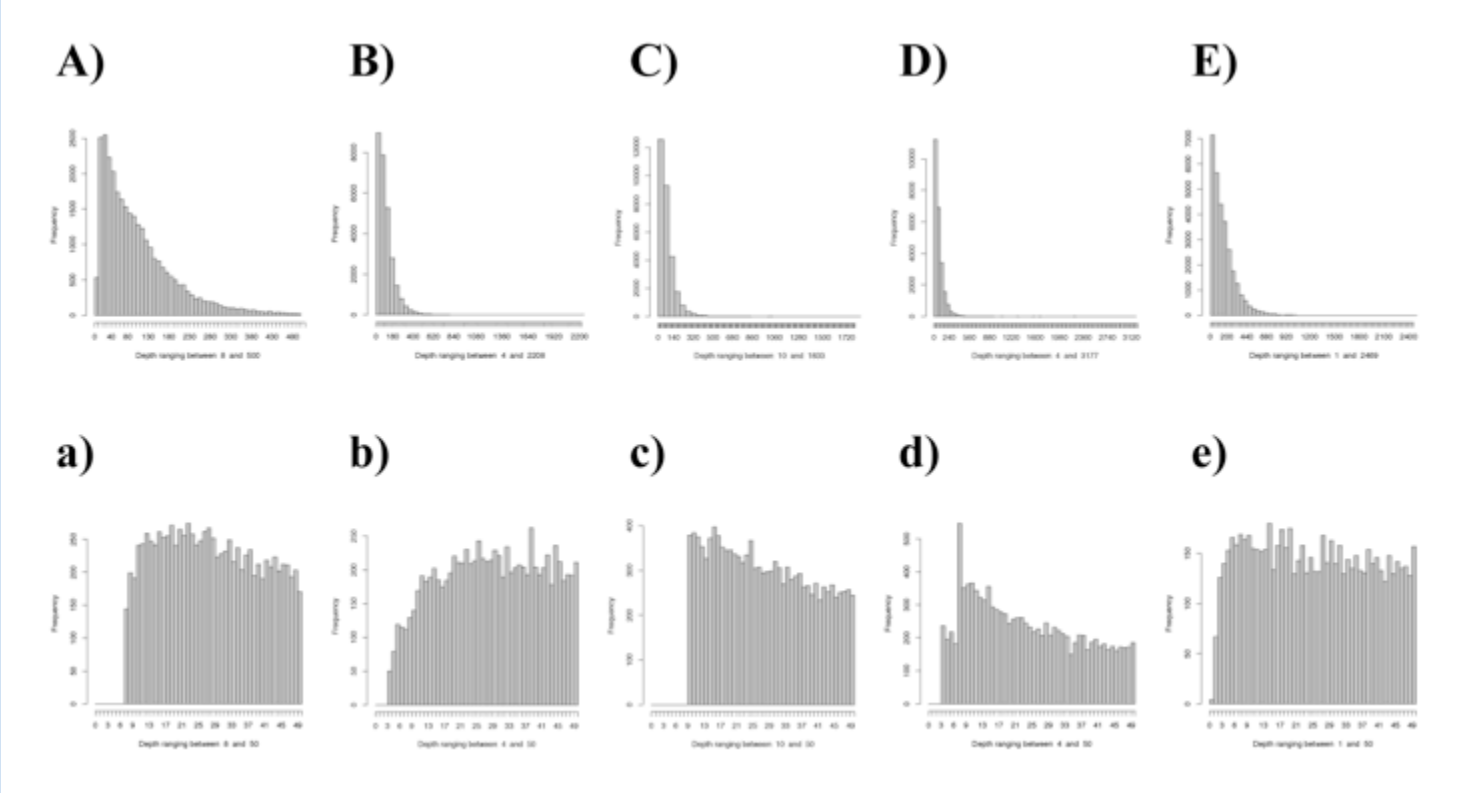


Validation of SNVs and Indels with an additional platform

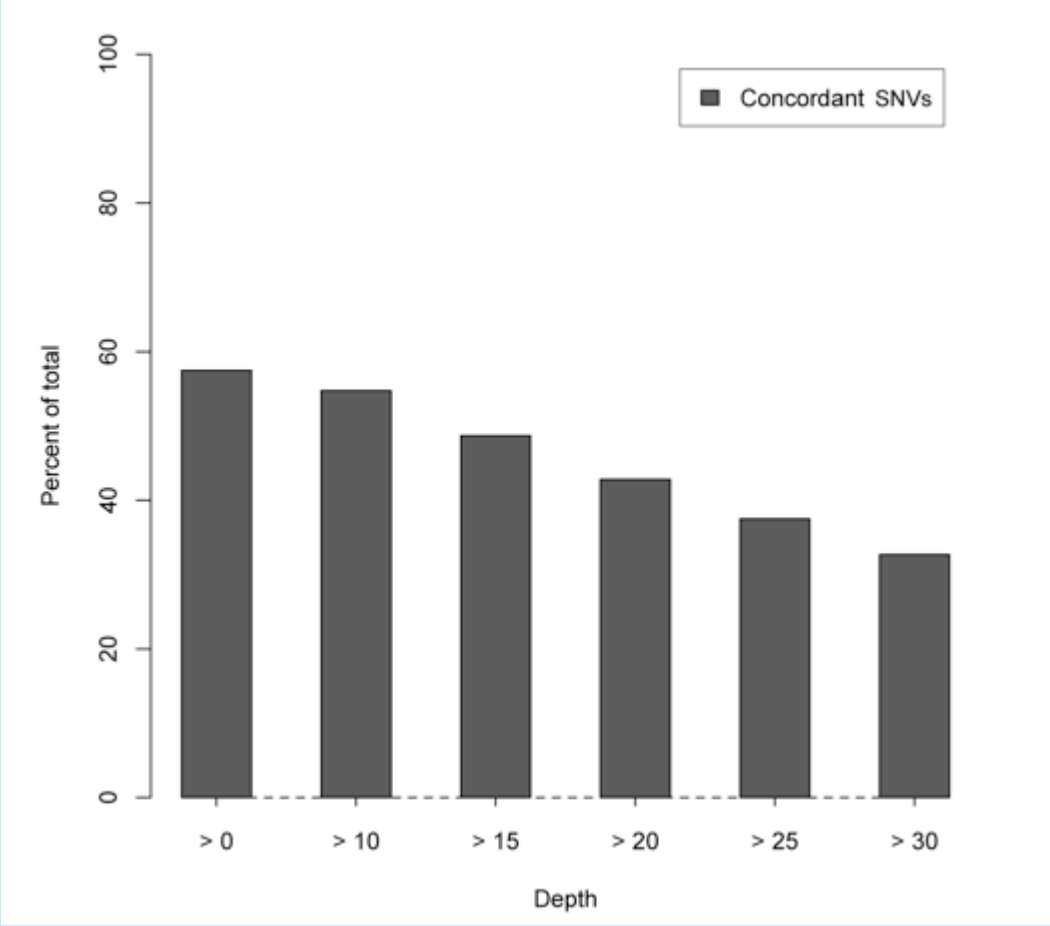




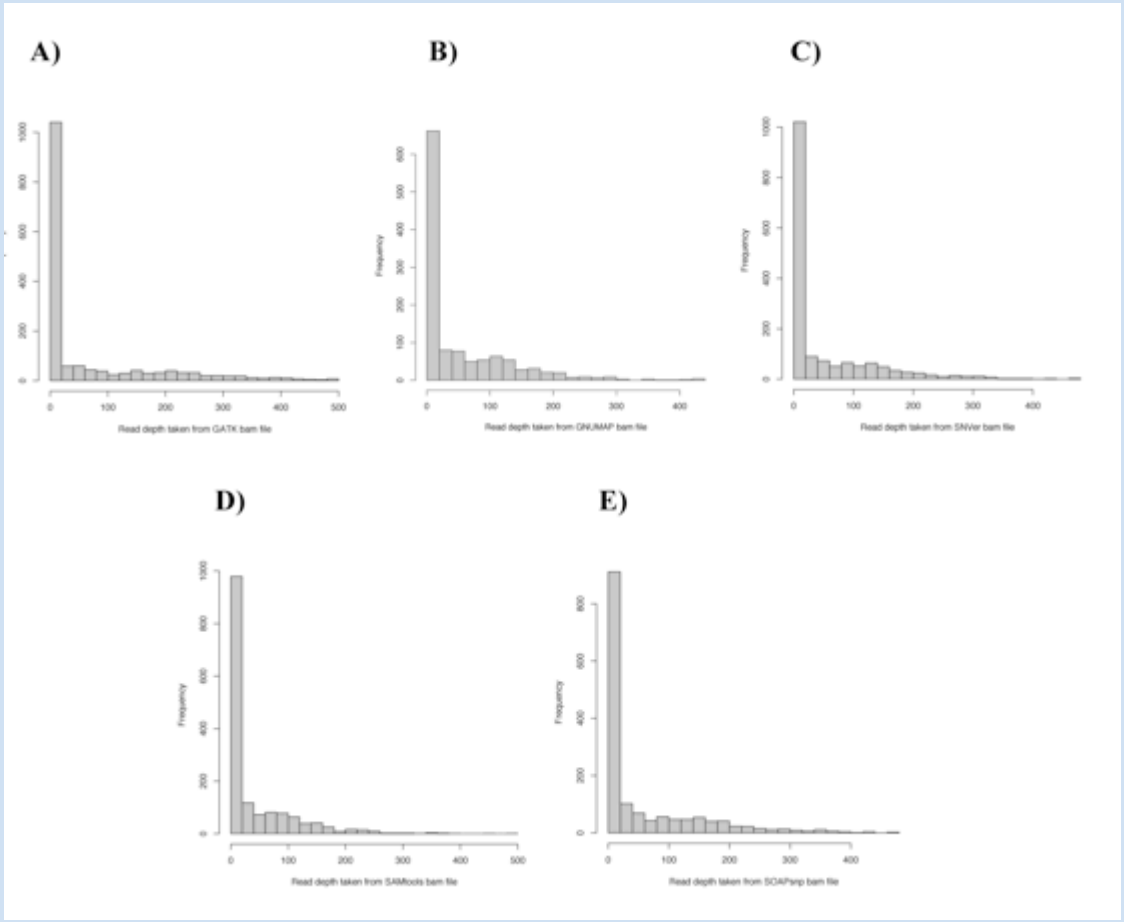
Read Depth



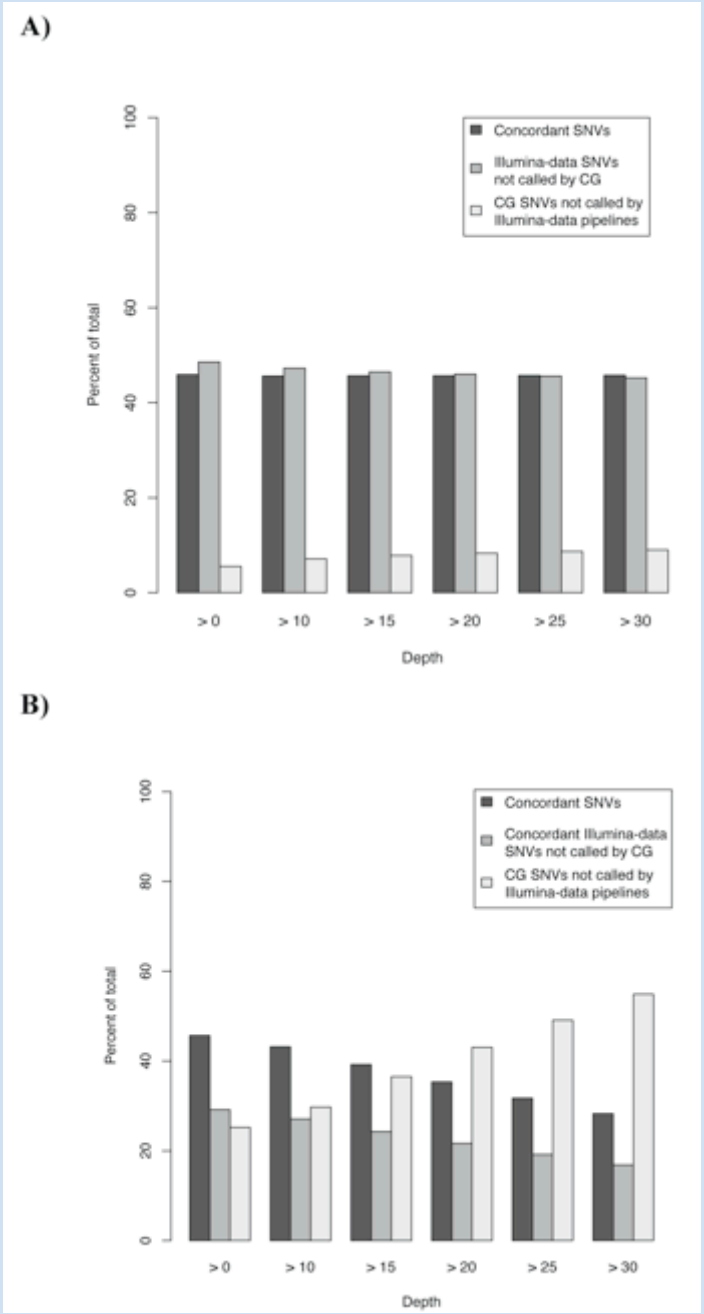
Additional file 2, Figure S3. Histograms of Illumina read depth taken from each pipeline's independently aligned BAM file at genomic coordinates of SNVs called by each of the 5 alignment and variant calling pipelines. **A)** SOAPsnp, **B)** SNVer, **C)** SAMTools, **D)** GNUMAP and **E)** GATK, respectively. Frequency of read depths for all SNVs (**A**, **B**, **C**, **D**, and **E**) as well as for SNVs having depths between 0 and 50 (**a**, **b**, **c**, **d**, and **e**) were plotted.

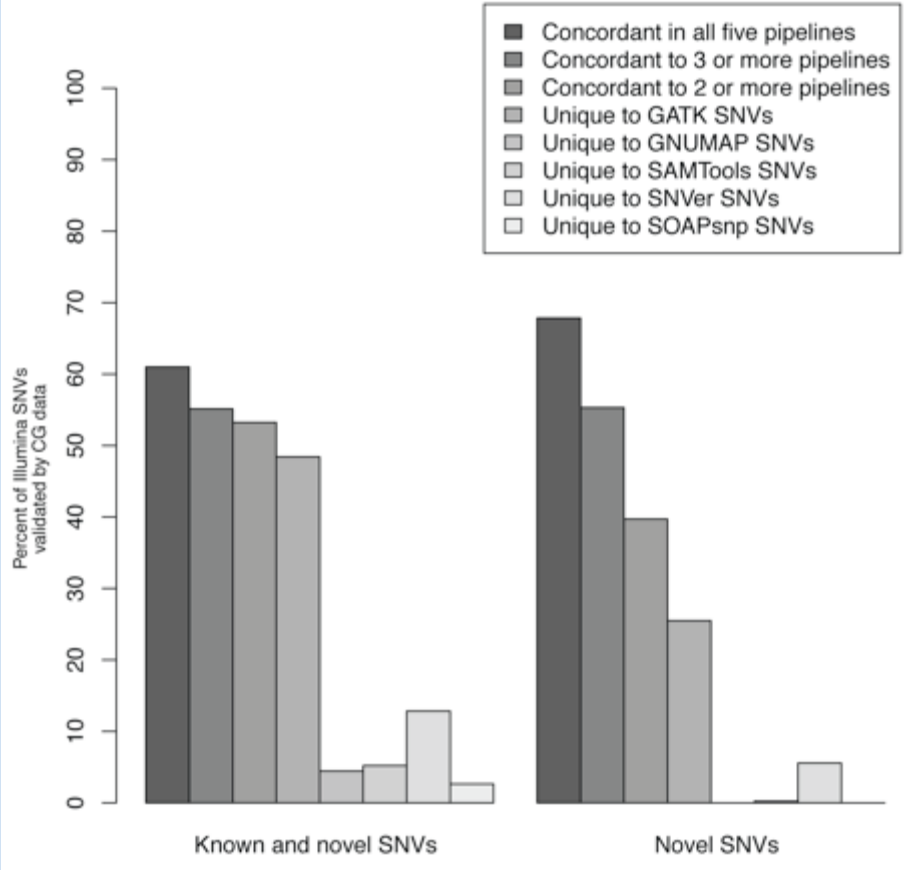


Additional file 2, Figure S4. SNV concordance for a single exome, "k8101-49685", between five alignment and variant detection pipelines: GATK, SOAPsnp, SNVer, SAMTools, and GNUMAP. Concordance between each pipeline was determined by matching the genomic coordinate as well as the base pair change and zygosity for each detected SNV. Concordance was measured at varying Illumina read depth threshold values in each independently aligned BAM file, ranging from >0 (no threshold) to >30 reads.

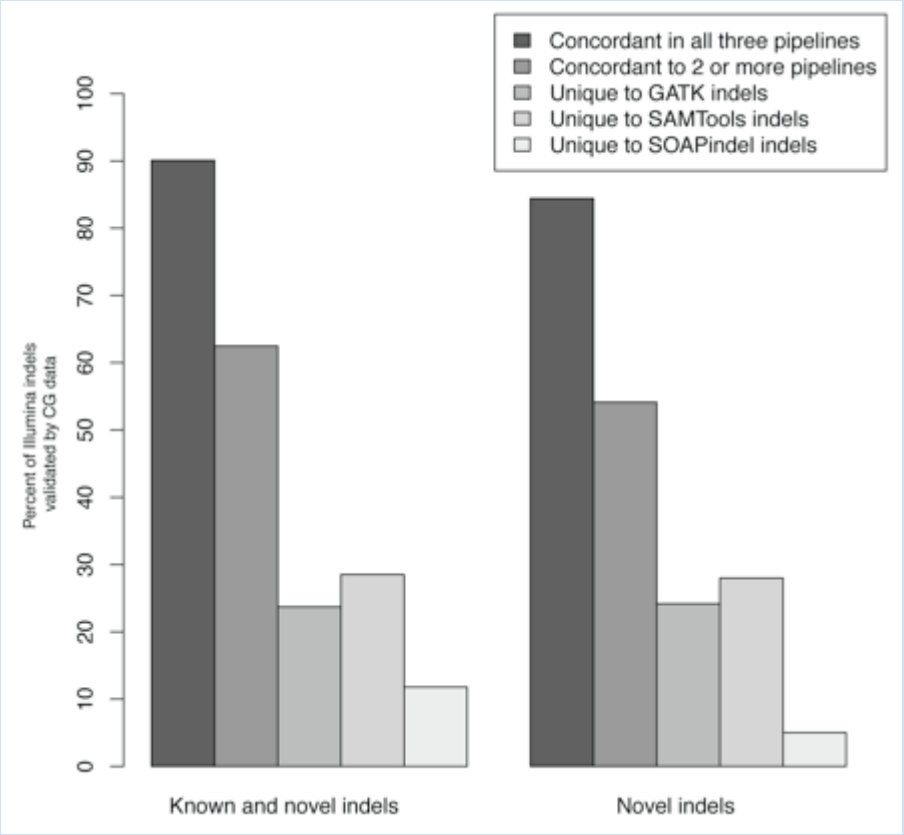


Additional file 2, Figure S5. Histograms of read depth taken from each of the five Illumina pipeline's independently aligned BAM file at genomic coordinates of SNVs that were found by Complete Genomics but not by any of the 5 Illumina pipelines: GATK, GNUMAP, SNVer, SAMTools and SOAPsnp, **A**, **B**, **C**, **D** and **E** respectively. All coordinates fell within the range of the Agilent SureSelect v.2 exons.


A)
Concordant SNVs
Illumina-data SNVs not called by CG
CG SNVs not called by Illumina-data pipelines
Percent of total
0
20
40
60
80
100
> 0
> 10
> 15
> 20
> 25
> 30
Depth
B)
Concordant SNVs
Concordant Illumina-data SNVs not called by CG
CG SNVs not called by Illumina-data pipelines
Percent of total
0
20
40
60
80
100
> 0
> 10
> 15
> 20
> 25
> 30
Depth

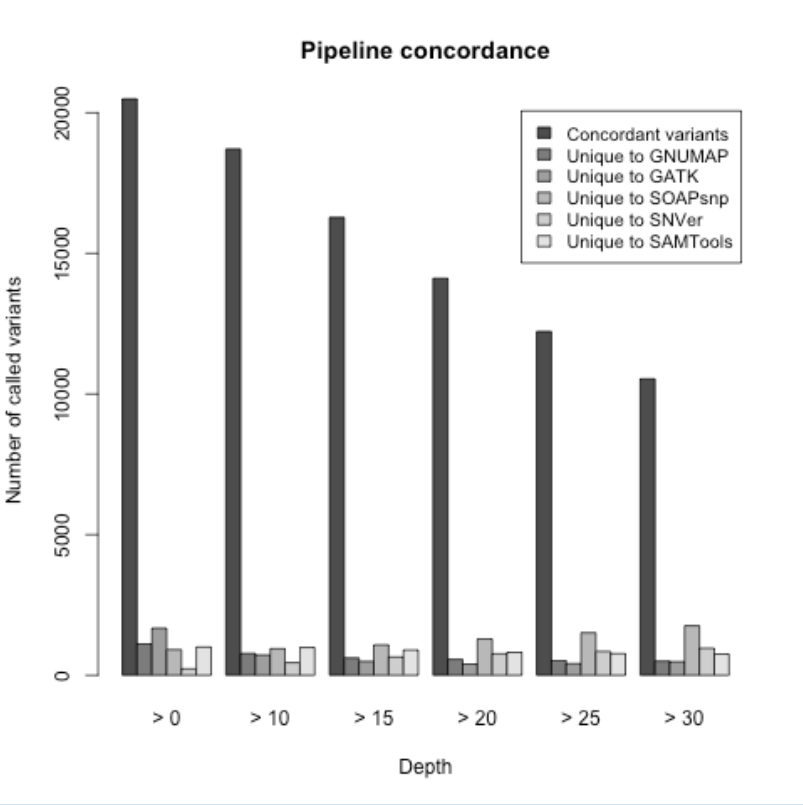


Additional file 2, Figure S8. SNVs called by each Illumina-data pipeline were cross-validated using SNVs called by Complete Genomics, an orthogonal sequencing technology, in sample "k8101-49685". The percentage of Illumina SNVs that were validated by CG sequencing was measured for variants having varying degrees of Illumina-data pipeline concordance. The same analysis was performed for variants that were considered novel (absent in dbSNP135).



Additional file 2, Figure S9. Indels called by each Illumina-data pipeline were cross-validated using indels called by Complete Genomics for sample "k8101-49685". The percentage of Illumina indels that were validated by CG sequencing was measured across varying degrees of Illumina pipeline concordance. The same analysis was done for novel indels (indels not found in dbSNP 135).

Comparing the concordance among the 5 pipelines used to analyze Illumina data, also stratified by read depth from >0 to >30 reads.



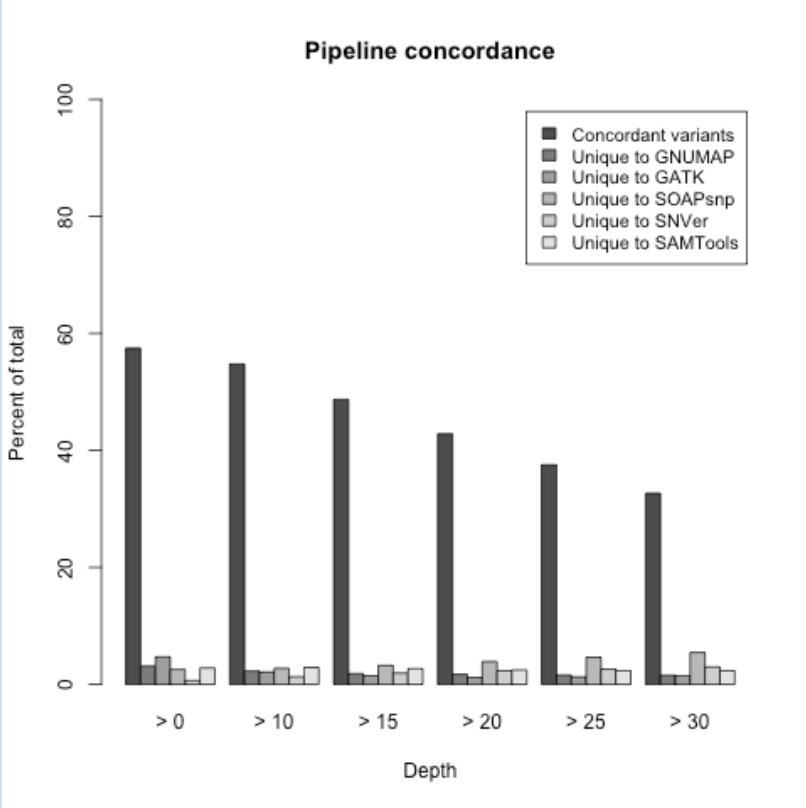


Table 2. Quality evaluation of variant detection using different variant-calling pipelines.

	Sensitivity		Specificity	
	Mean*	SD	Mean*	SD
SOAPsnp	94.68	2.26	99.79	0.03
GATK1.5	95.34	1.16	99.72	0.08
SNVer	92.33	4.40	99.78	0.04
GNUMAP	86.60	3.23	99.64	0.06
SAMtools	94.47	4.22	99.59	0.16
Any pipeline	97.67	1.20	99.62	0.11
≥2 pipelines*	96.64	2.28	99.69	0.07
≥3 pipelines*	95.62	3.13	99.73	0.05
≥4 pipelines*	92.60	3.40	99.82	0.04
5 pipelines*	80.58	5.26	99.87	0.01

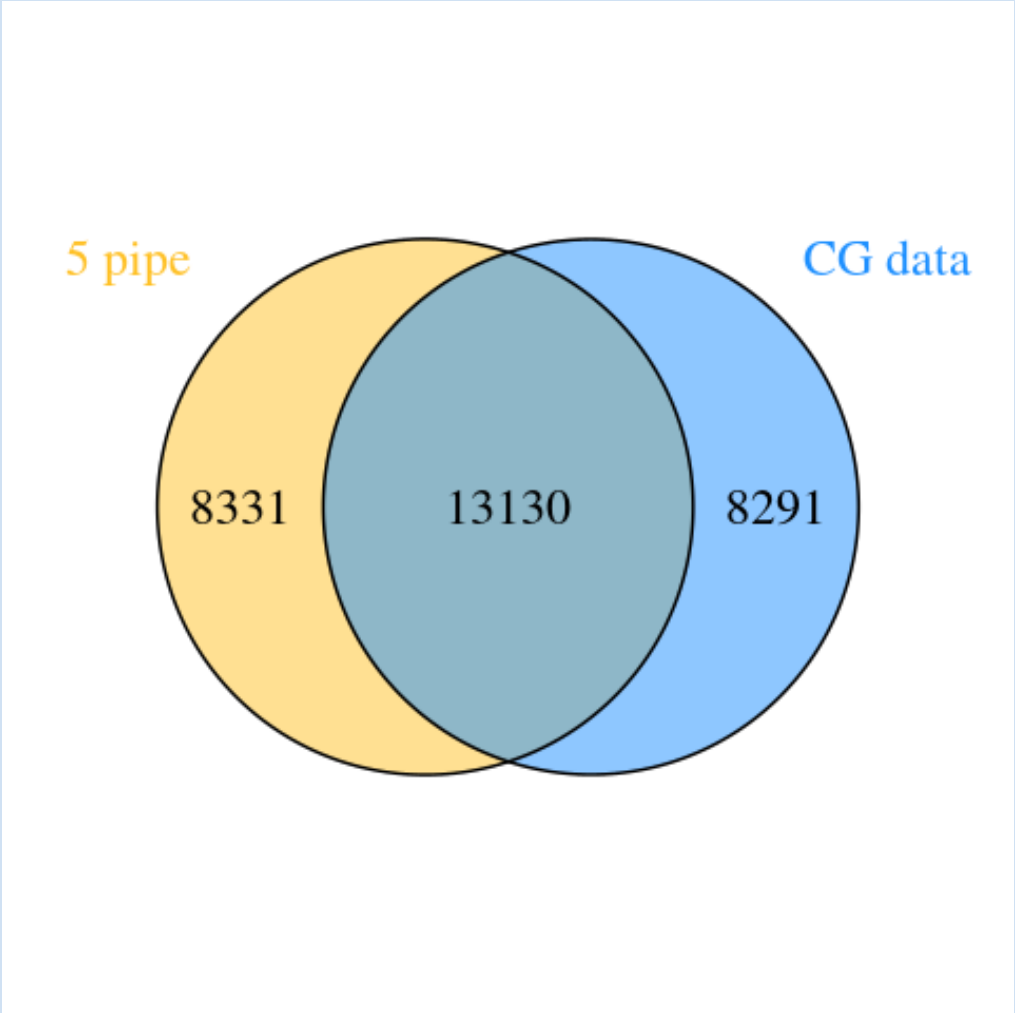
*Intersection of variants contained in the number of pipelines specified. Sensitivity and specificity was calculated for each pipeline by comparing Illumina Human610-Quad version 1 SNP arrays with exome-capture sequencing results, based on the four samples whose genotyping data was available.

Table S1. Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips.

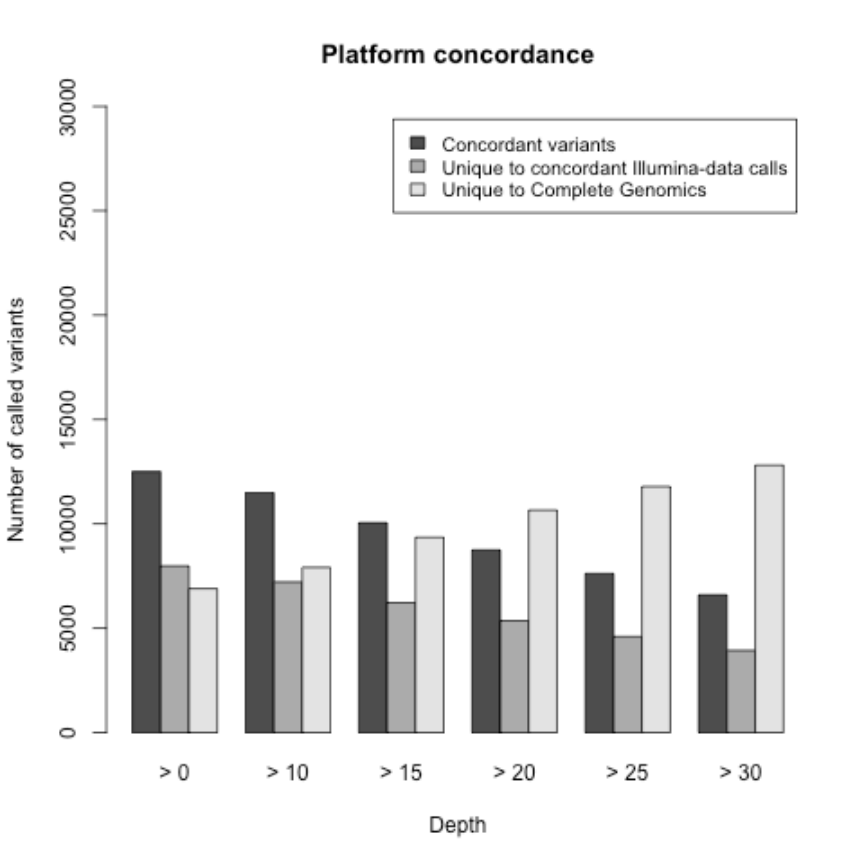
Sample	Software	Compared Sites	Concordance Sites	Concordance rate
Mother-1	SOAPsnp	6088	6074	99.77%
	GATK 1.5	6249	6224	99.60%
	SNVer	5723	5708	99.74%
	GNUMAP	5458	5434	99.56%
	SAMTools	5885	5848	99.37%
Son-1	SOAPsnp	6366	6353	99.80%
	GATK 1.5	6341	6323	99.72%
	SNVer	6255	6239	99.74%
	GNUMAP	5850	5828	99.62%
	SAMTools	6383	6362	99.67%
Son-2	SOAPsnp	6412	6401	99.83%
	GATK 1.5	6426	6413	99.80%
	SNVer	6336	6325	99.83%
	GNUMAP	5906	5889	99.71%
	SAMTools	6477	6450	99.58%
Father-1	SOAPsnp	6247	6238	99.86%
	GATK 1.5	6304	6288	99.75%
	SNVer	6205	6192	99.79%
	GNUMAP	5805	5786	99.67%
	SAMTools	6344	6327	99.73%

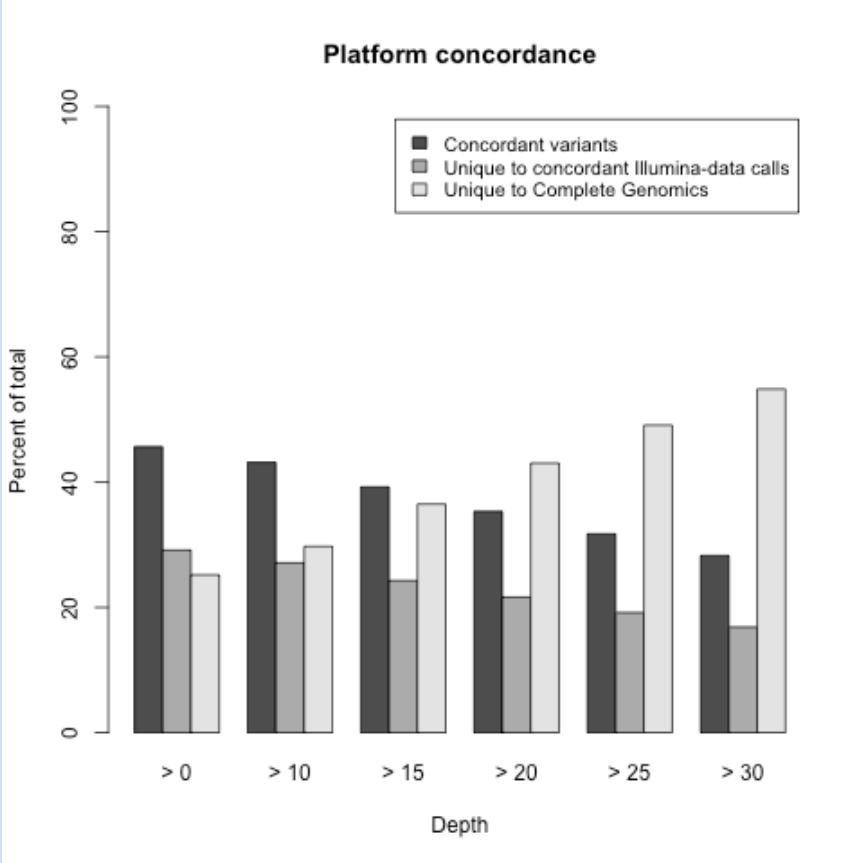
All pipelines are very good with identifying already known, common SNPs.

Taking SNVs concordant in 5 Illumina pipelines, and comparing to SNVs in Complete Genomics Data from same sample

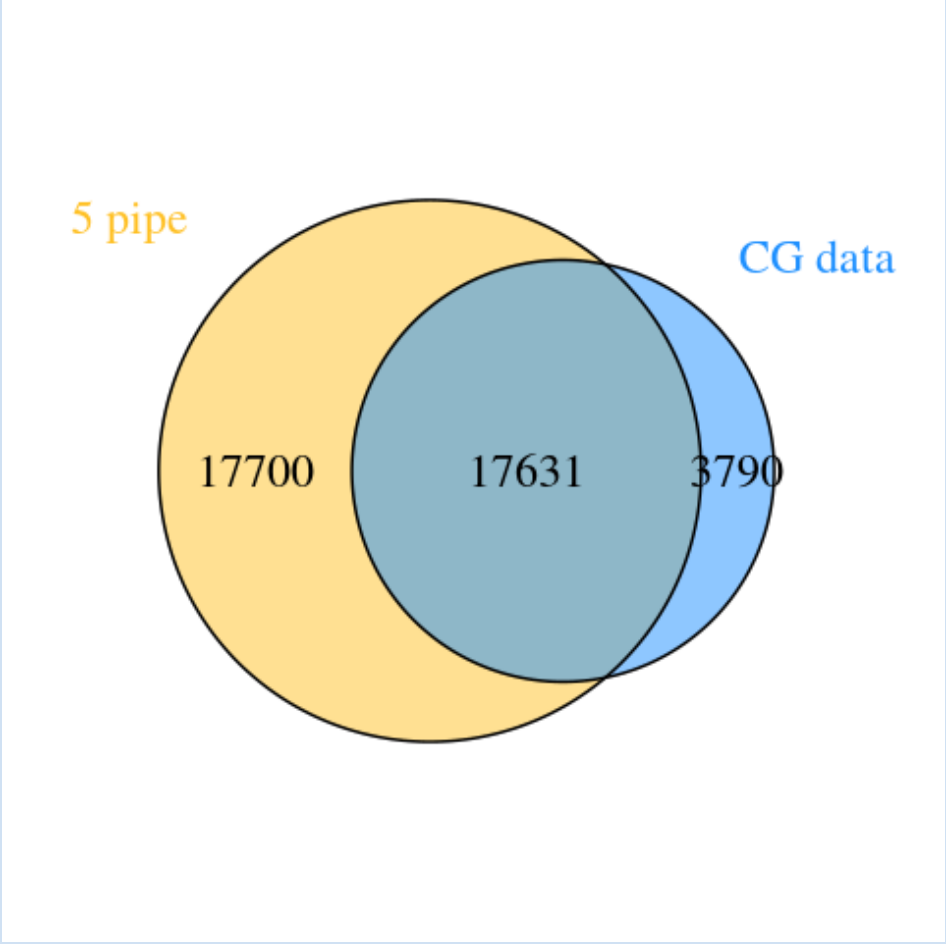


Taking SNVs concordant in 5 Illumina pipelines as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample

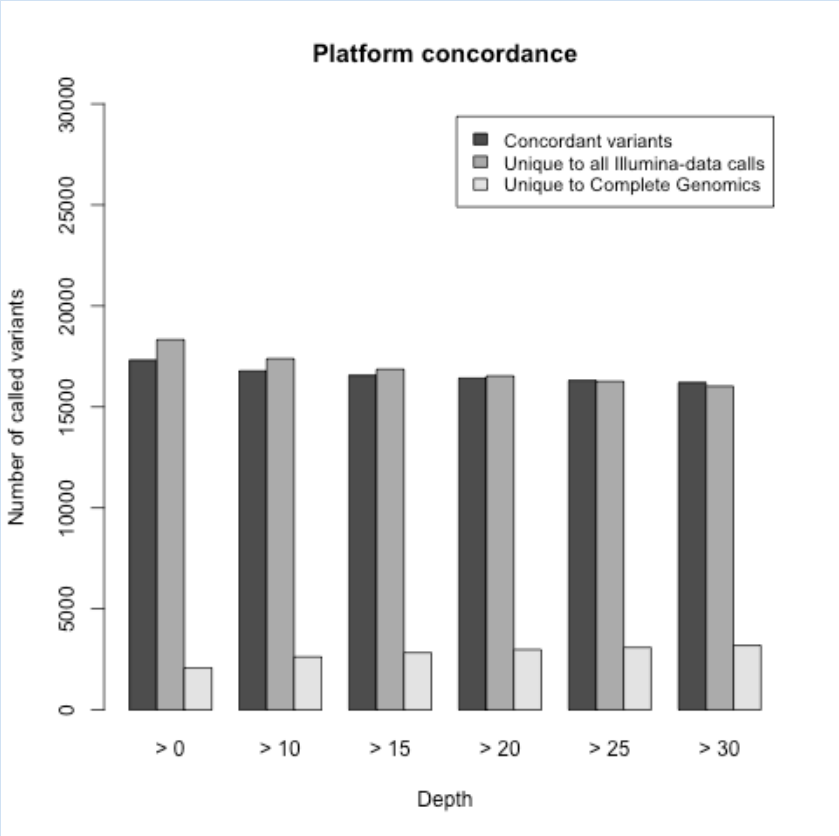


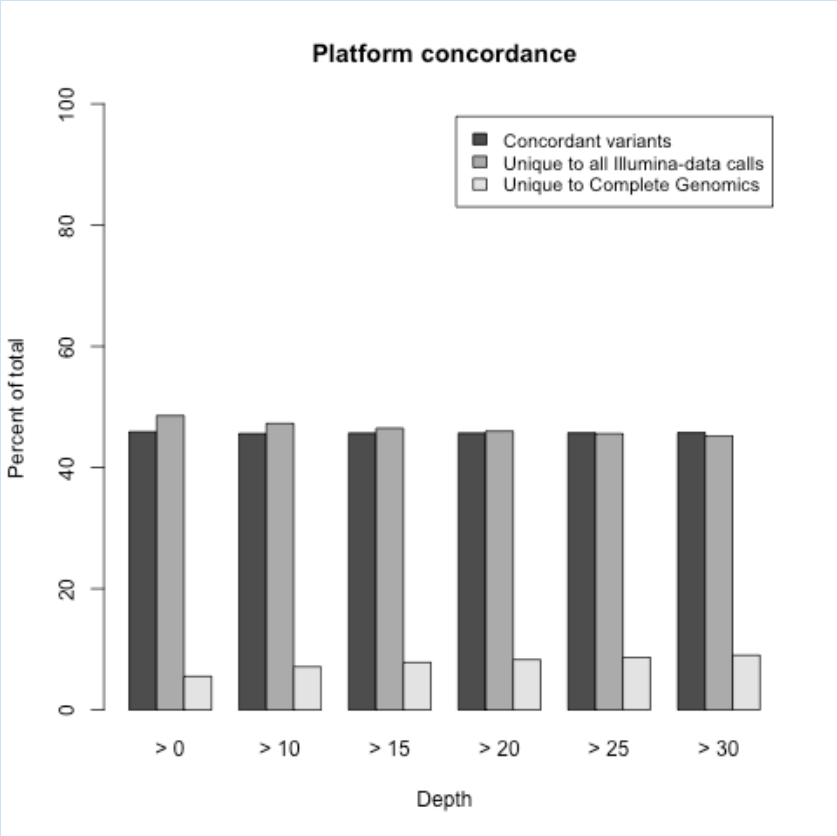


Taking SNVs found by ALL 5 Illumina pipelines (Union), and comparing to SNVs in Complete Genomics Data from same sample

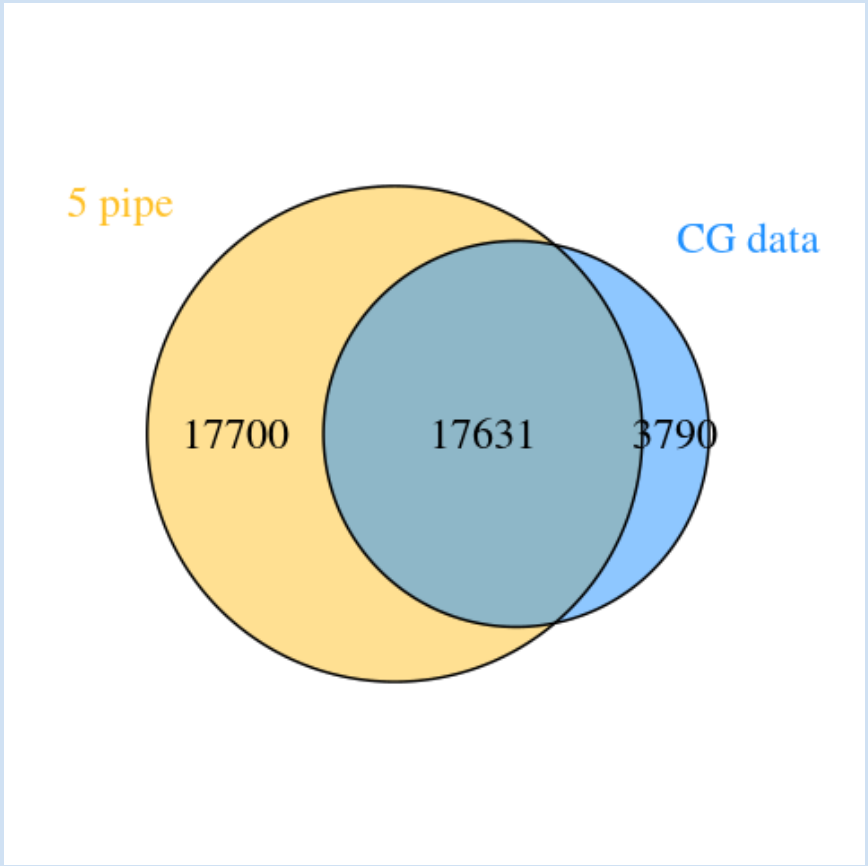


Taking the UNION of all SNVs called by Illumina pipelines, as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample

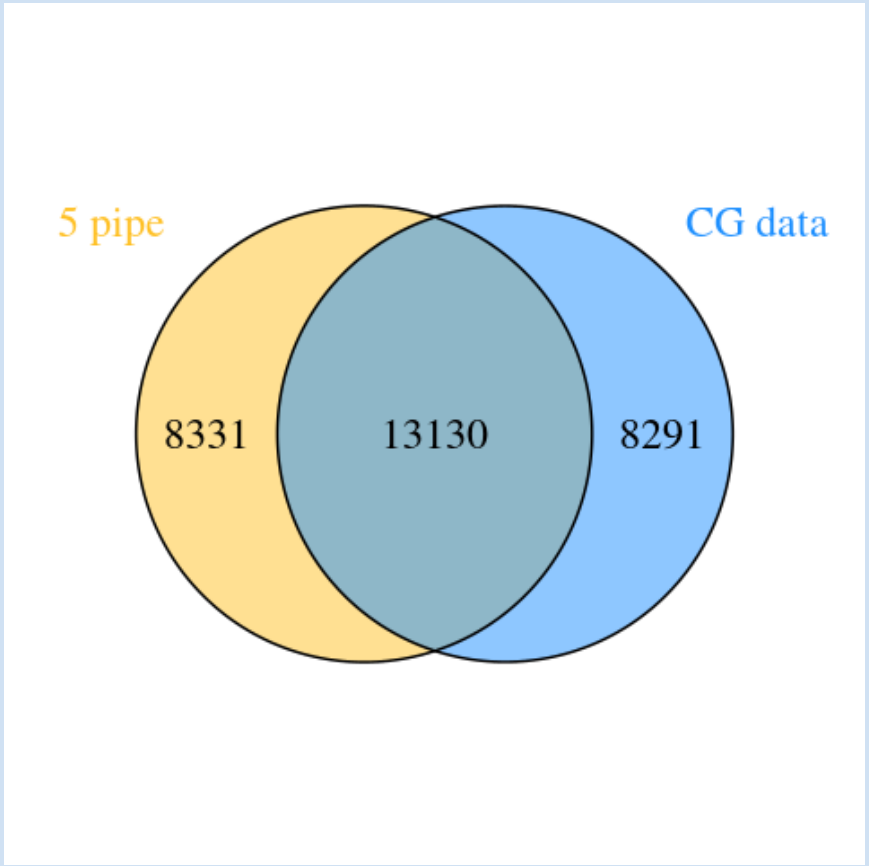




Comparing the UNION versus the CONCORDANCE of 5 pipelines to the Complete Genomics Data

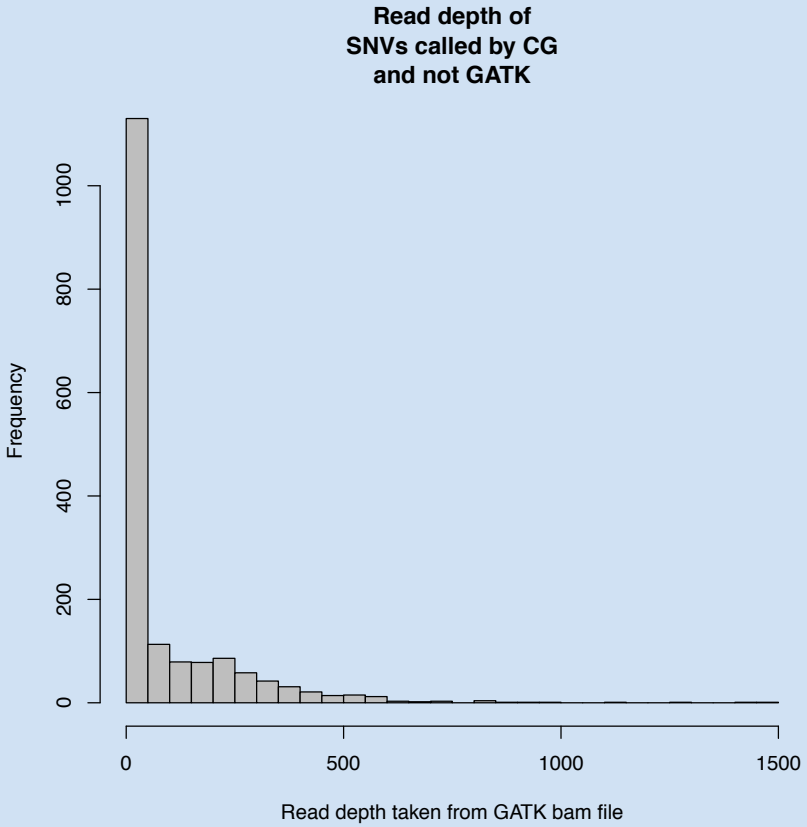


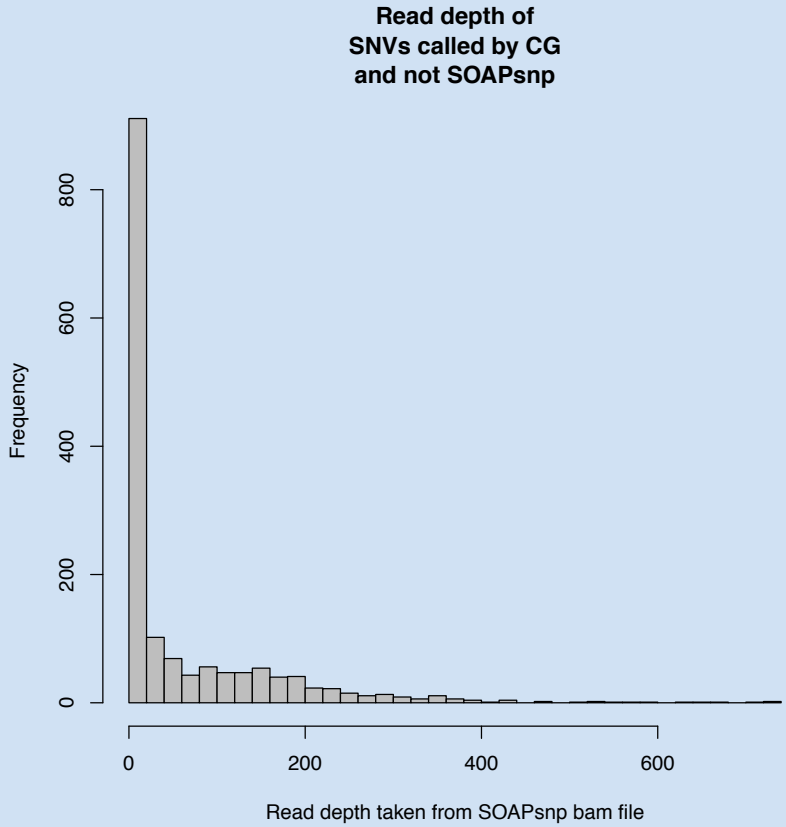
Union of Illumina variants



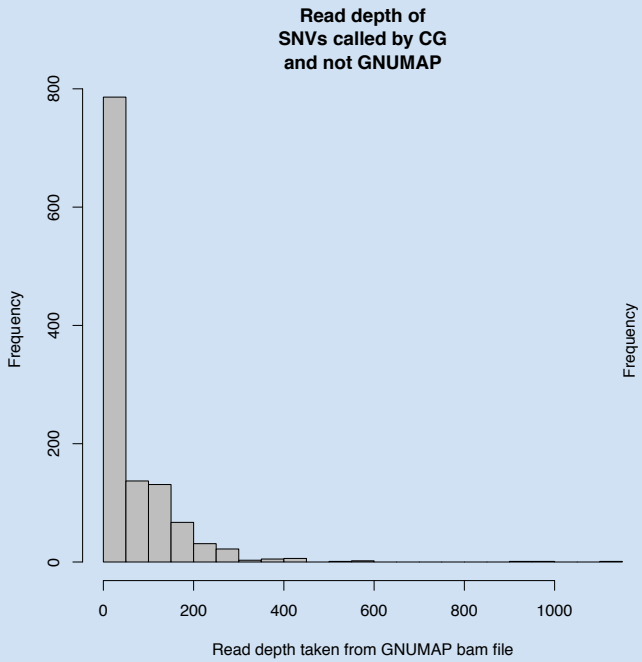
Concordant Illumina variants

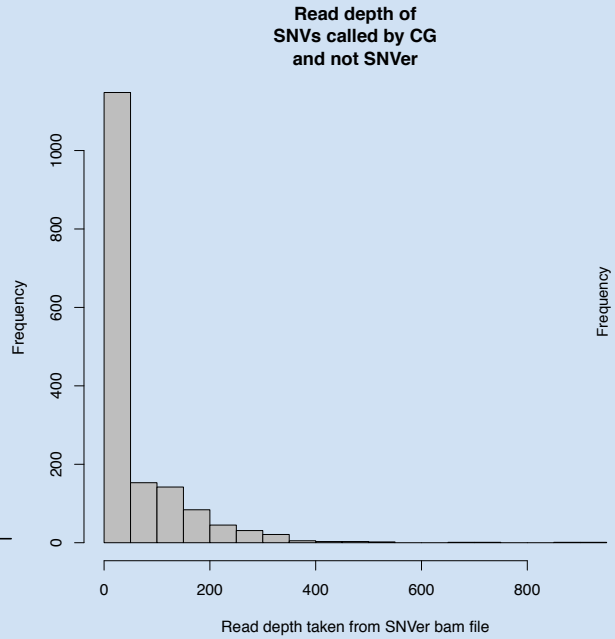
Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GATK or SOAP pipelines

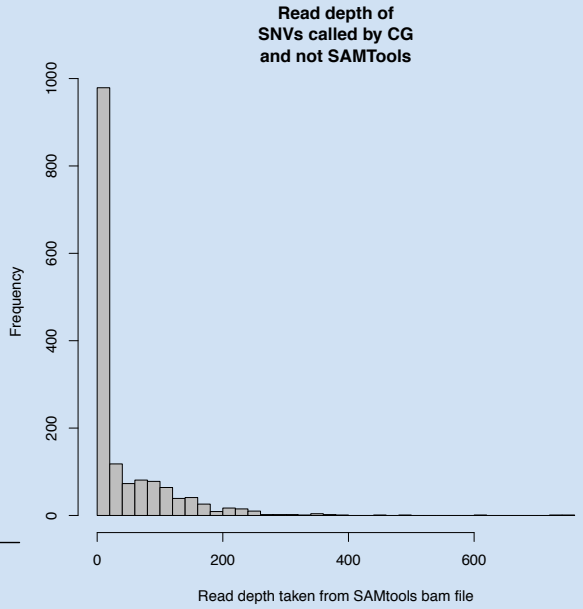




Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GNUMAP, SNVer or SamTools pipelines







Genomic Dark Matter, cont....

- That means that unlike typical false negatives, increasing coverage will not help identify mutations in low GMS regions, even with 0% sequencing error.
- Instead this is because the SNP-calling algorithms use the mapping quality scores to filter out unreliable mapping assignments, and low GMS regions have low mapping quality score (by definition). Thus even though many reads may sample these variations, the mapping algorithms cannot ever reliably map to them.
- Since about 14% of the genome has low GMS value with typical sequencing parameters, it is expected that about 14% of all variations of all resequencing studies will not be detected.
- To demonstrate this effect, we characterised the SNP variants identified by the 1000 genomes pilot project, and found that 99.99% of the SNPs reported were in high GMS regions of the genome, and in fact 99.95% had GMS over 90.

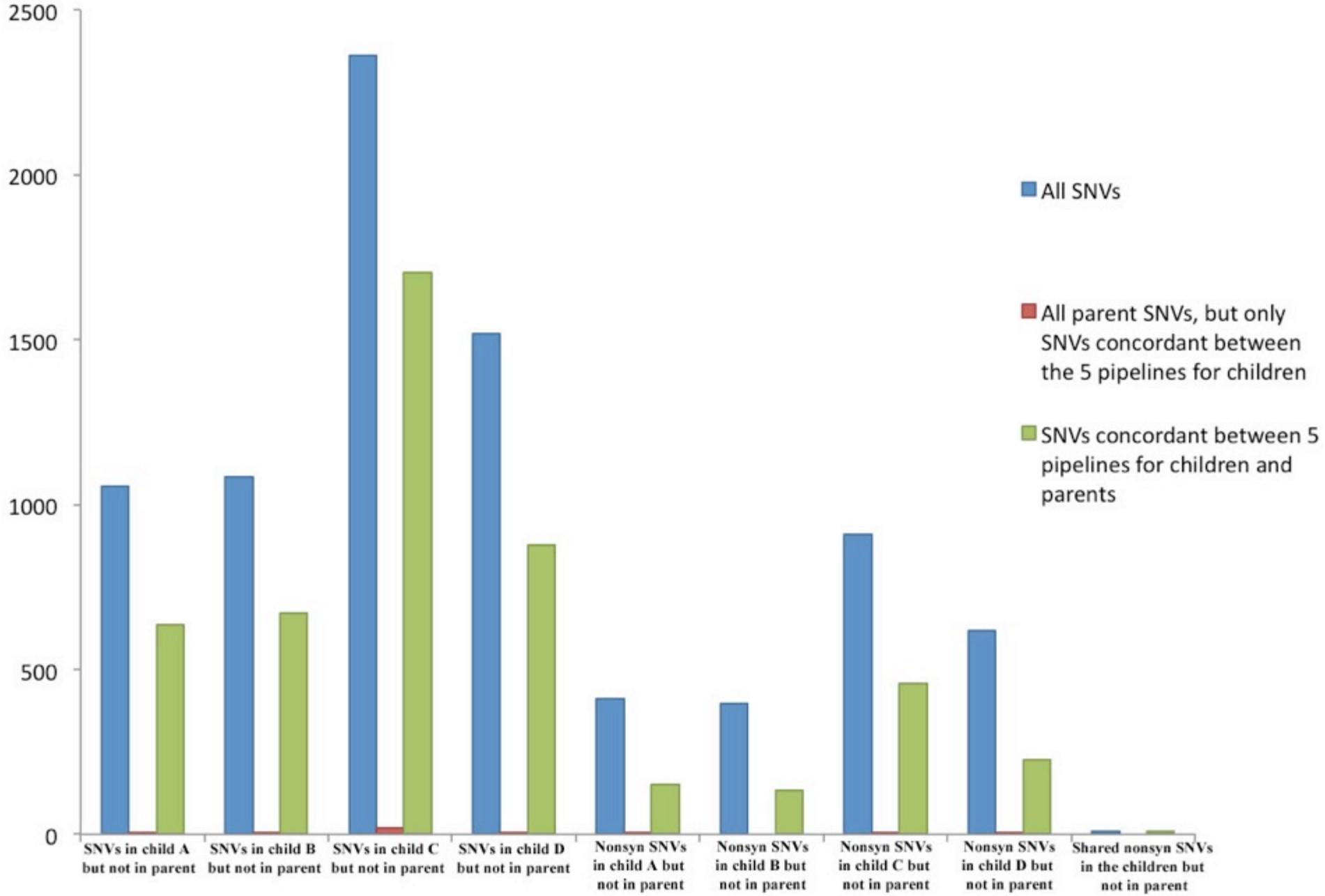

2500
2000
1500
1000
500
0
All SNVs
All parent SNVs, but only SNVs concordant between the 5 pipelines for children
SNVs concordant between 5 pipelines for children and parents
SNVs in child A but not in parent
SNVs in child B but not in parent
SNVs in child C but not in parent
SNVs in child D but not in parent
Nonsyn SNVs in child A but not in parent
Nonsyn SNVs in child B but not in parent
Nonsyn SNVs in child C but not in parent
Nonsyn SNVs in child D but not in parent
Shared nonsyn SNVs in the children but not in parent

Figure 1. Mean single-nucleotide variants (SNV) concordance over 15 exomes between five alignment and variant-calling pipelines. The alignment method used, followed by the SNV variant calling algorithm is annotated here in shorthand: BWA-GATK, SOAP-Align-SOAPsnp, BWA-SNVer, BWA-SAMtools, and GNUMAP-GNUMAP. **(A)** Mean SNV concordance between each pipeline was determined by matching the genomic coordinate as well as the base-pair change and zygosity for each detected SNV. **(B)** The same analysis as in (A) but filtered to include only SNVs already found in dbSNP135. **(C)** The same analysis as in (A), but filtered to include novel SNVs (that is, SNVs not found in dbSNP135).

Pipelines Used on Same Set of Seq Data by Different Analysts, using Hg19 Reference Genome

1) BWA - Sam format to Bam format - Picard to remove duplicates - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper.

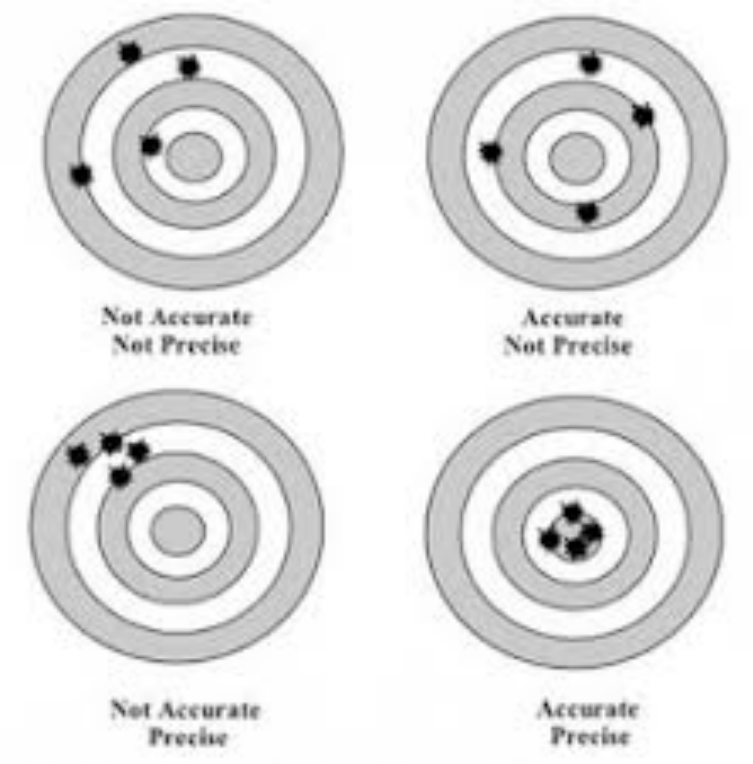
2) BWA - Sam format to Bam format-Picard to remove duplicates - **SamTools** version 0.1.18 to generate genotype calls -- The "mpileup" command in SamTools were used for identify SNPs and indels.

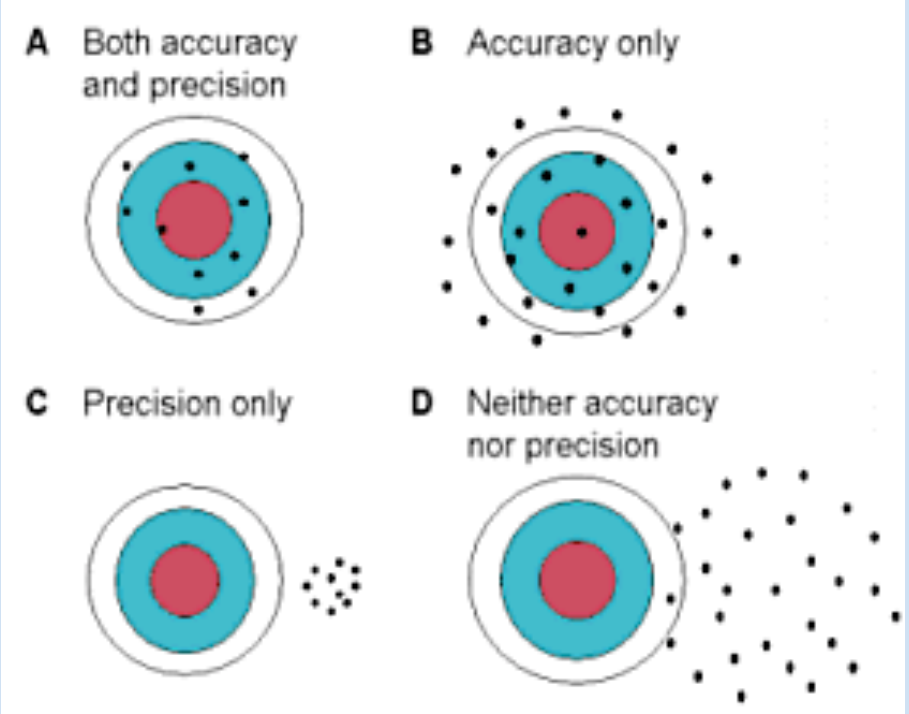
3) SOAP-Align – SOAPsnp – then BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph)

4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion)

5) BWA - Sam format to Bam format - Picard to remove duplicates - **SNVer**

6) BWA - Sam format to Bam format - Picard to remove duplicates - **SCALPEL**


Not Accurate
Not Precise
Accurate
Not Precise
Not Accurate
Precise
Accurate
Precise


A Both accuracy and precision
B Accuracy only
C Precision only
D Neither accuracy nor precision

Pipelines Used on Same Set of Seq Data by Different Analysts, using Hg19 Reference Genome

1) BWA - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

2) BWA - **SamTools** version 0.1.18 to generate genotype calls -- The "mpileup" command in SamTools was used for identify SNVs and indels.

3) SOAP-Align – SOAPsnp for SNVs– and BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph) for indels.

4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion), for SNVs only.

5) BWA - Sam format to Bam format - Picard to remove duplicates **– SNVer** , for SNVs only

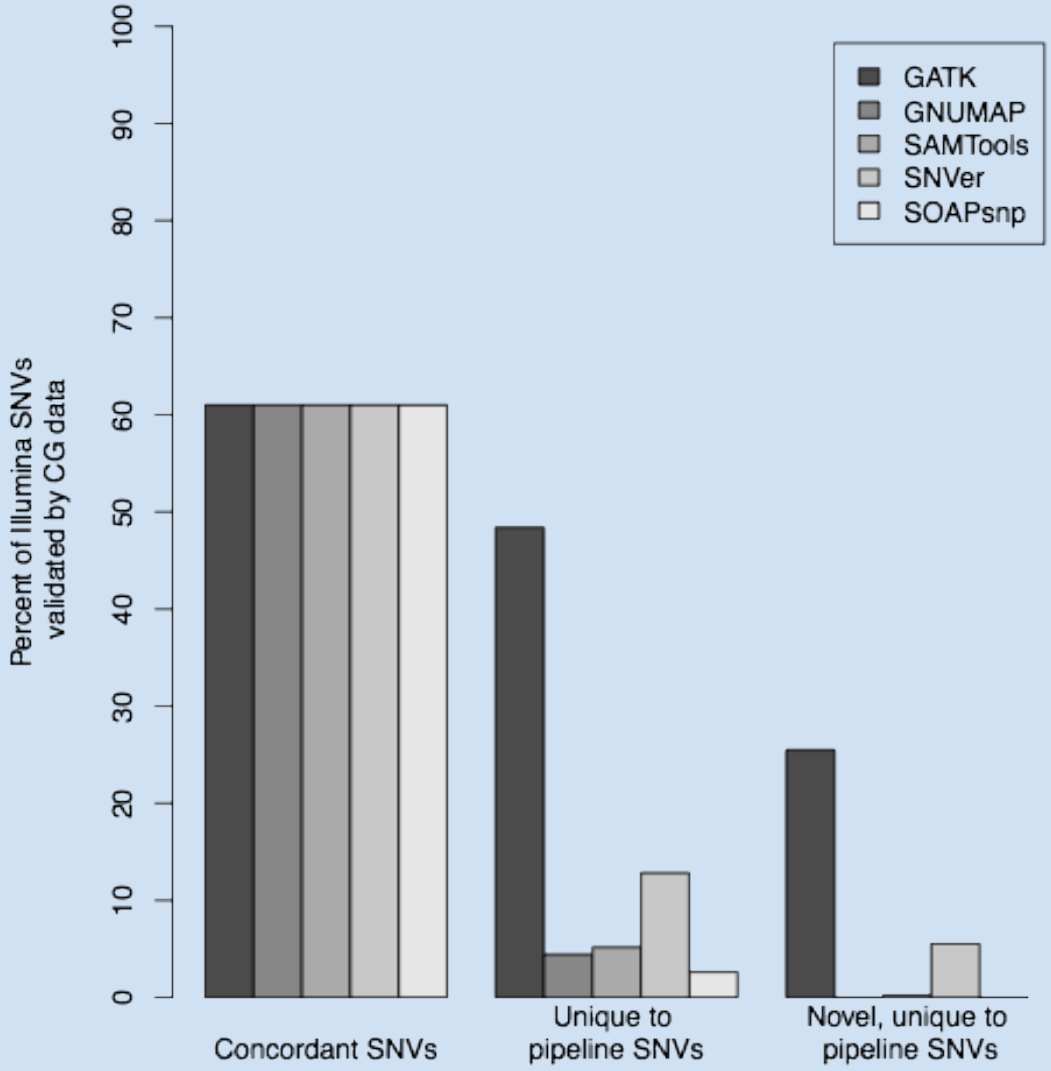
	All SNVs, both for parents and children, were considered	All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children	SNVs concordant between 5 pipelines for children and parents
Number of SNVs found in child A but not in parents	1057	2	637
Number of SNVs found in child B but not in parents	1084	1	672
Number of SNVs found in child C but not in parents	2363	20	1703
Number of SNVs found in child D but not in parents	1518	5	876
Number of nonsyn SNVs in child A but not in parents	411	1	150
Number of nonsyn SNVs in child B but not in parents	396	0	135
Number of nonsyn SNVs in child C but not in parents	911	6	459
Number of nonsyn SNVs in child D but not in parents	619	3	225
Number of shared nonsyn SNVs in the children, but not in parents	8	0	9

Optimizing pipeline based on literature value of ~1 true de novo protein-altering mutation per exome

	All SNVs, both for parents and children, were considered	All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children	SNVs concordant between 5 pipelines for children and parents
Number of SNVs found in child A but not in parents	1308	186	1795
Number of SNVs found in child B but not in parents	1332	161	1762
Number of nonsyn SNVs in child A but not in parents	381	52	420
Number of nonsyn SNVs in child B but not in parents	392	42	394
Number of shared nonsyn SNVs in the children, but not in parents	98	14	171

The result is that using all of the detected SNVs for both parents and children should minimize the false negative rate but similarly show a relatively high false positive rate. Using all of the SNVs detected for parents but only the SNVs concordant among the five pipelines shows mutation rates similar to those reported by the literature and is expected to have moderate false positive rates and moderate false negative rates. Using only the SNVs concordant among the 5 different pipelines for both parents and children should minimize the false positive rate but similarly show a relatively high false negative rate.

Much Higher Validation of the Concordantly Called SNVs (by the CG data)



CSH Cold Spring Harbor Laboratory

STANLEY INSTITUTE FOR COGNITIVE GENOMICS

THE
STANLEY MEDICAL
RESEARCH INSTITUTE

NIMH
National Institute
of Mental Health